RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

PROCEEDINGS OF THE FOURTH INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE

Volume 1

BGRS' 2004
Novosibirsk, Russia
July 25 - 30, 2004

IC&G, Novosibirsk, 2004

International Program Committee

Nikolay Kolchanov, Institute of Cytology and Genetics, Novosibirsk, Russia
(Chairman of the Conference)

Ralf Hofestaedt University of Bielefeld, Germany (Co-Chairman of the Conference)

Dagmara Furman, Institute of Cytology and Genetics, Novosibirsk,
(Conference Scientific Secretary)

Jurgen Borlak, Center of Drug Research and Medical Biotechnology, Fraunhofer Institute of Toxicology and Experimental Medicine, Hannover, Germany

Philipp Bucher, Swiss Institute for Experimental Cancer Research, Switzerland

Gennady Erokhin, Ugra Research Institute of Information Technologies,

Khanty-Mansiysk, Russia

Jim Fickett, AstraZeneca, Boston, USA

Mikhail Gelfand, GosNIIGenetika, Moscow, Russia

Sergey Goncharov, Sobolev Institute of Mathematics, Novosibirsk, Russia

Igor Goryanin, GlaxoSmithKline, UK

Charlie Hodgman, GlaxoSmithKline, UK

Lev Kisselev, Engelhardt Institute of Molecular Biology, Moscow, Russia

Victor Malyshkin, Institute of Computational Mathematics and Mathematical Geophysics, Novosibirsk, Russia

Luciano Milanesi, National Research Council - Institute of Biomedical Technology, Italy

Eric Mjolsness, Institute for Genomics and Bioinformatics, University of California, Irvine, USA

Nilkolay Podkolodny Institute of Cytology and Genetics, Novosibirsk, Russia

Akinori Sarai, Kyushu Institute of Technology (KIT), Iizuka, JAPAN

Rustem Tchuraev, Institute of Biology, Ufa Scientific Centre RAS, Ufa, Russia

Denis Thieffry, ESIL-GBMA, Universite de la Mediterranee, Marseille, France

Masaru Tomita, Institute for Advanced Biosciences, Keio University, Japan

Alexander Vershinin, Institute of Cytology and Genetics, Novosibirsk, Russia

Edgar Wingender, UKG, University of Goettingen, Goettingen, Germany

Eugene Zabarovsky, Karolinska Institute, Stockholm, Sweden

Lev Zhivotovsky, Institute of General Genetics, Moscow, Russia

Local Organizing Committee

Sergey Lavryushev, Institute of Cytology and Genetics, Novosibirsk (Chairperson)

Anatoly Kushnir, Institute of Cytology and Genetics, Novosibirsk

Natalia Sournina, Institute of Cytology and Genetics, Novosibirsk

Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk

Katerina Denisova, Institute of Cytology and Genetics, Novosibirsk

Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk

Yuri Orlov, Institute of Cytology and Genetics, Novosibirsk

The information about the Conference BGRS' 2004 can be found at http://www.bionet.nsc.ru/meeting/bgrs2004/

Organizers

Institute of Cytology and Genetics,
Siberian Branch of the Russian Academy of Sciences

Siberian Branch of the Russian Academy of Sciences

All - Russian Society for Geneticists and Breeders

Ugra Research Institute of Information Technologies

INTAS

Sponsors

Russian Foundation for Basic Research

AstraZeneca, Boston, USA

Information Sponsors

Biophysics (Russian)

In Silico Biology

Contents

COMPUTATIONAL STRUCTURAL AND FUNCTIONAL GENOMICS

COMPUTATIONAL STRUCTURAL AND FUNCTIONAL PROTEOMICS

Introduction

Two volumes of Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure – BGRS' 2004 (Akademgorodok, Novosibirsk, Russia, July 25-30, 2004) incorporate about 180 peer-reviewed publications (extended abstracts or short papers) devoted to the actual problems in bioinformatics of genome regulation and structure.

The Conference BGRS'2004 is organized by the Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. BGRS'2004 is the fourth in the series. It will continue the traditions of the previous conferences, BGRS'1998, BGRS'2000 and BGRS'2002, which were held in Novosibirsk in August 1998, 2000, and 2002, respectively.

BGRS'2004 provides a general forum for disseminating and facilitating the latest developments in bioinformatics in molecular biology. BGRS'2004 is a multidisciplinary conference. Its scope includes the development and application of advanced methods of computational and theoretical analysis for structure-function genome organization, proteomics, evolutionary and system biology. The scientists with an interest in bioinformatics, mathematical, theoretical or computational biology will attend the meeting. The event addresses the latest research in these fields, and will be a great opportunity for attendees to showcase their works.

Except researchers dealing with *in silico* approaches, the scientists involved in experimental research and interested in broad using theoretical and/or computational methods in their practice traditionally participate in the work of the conference. Thus, the conference creates an interface between experimental and computer-assisted researches in the fields of genomics, transcriptomics, proteomics, structural and systemic biology, as well as for contributing to promotion of computational biology to experimental research.

The post-genome era in biology is characterized by sharp increase in research scale in the fields of transcriptomics, proteomics, and systemic biology (gene interaction, gene network functioning, signal transduction pathways, networks of protein-protein interactions, etc.) without loosing the fundamental interest to studying structural genome organization.

The structure and regulation of genome are the counterparts of life at molecular level; that is why understanding of fundamental principles of regulatory genomic machinery is impossible unless genome structural organization is known, and *vice versa*.

The huge volume of experimental data that has been acquired on genome structure, functioning and gene expression regulation demonstrate the blistering growth. Onrush of the volumes of experimental data is observed in the recent years due to the fact that genome deciphering became a technical routine task produced at high speed. Unprecedented large bulk of experimental data emerge under studying of molecular-genetic systems and processes with application of microarray analysis technique. Unwrapping of large scale studying in proteomics is accompanied by accumulation of very large information pools on primary and spatial structures and functioning patterns. That is why development of informational-computational technologies of novel generation is a challenging problem of contemporary bioinformatics. Bioinformatics has entered that very phase of development, when decisions of the challenging problems determine the realization of large-scale experimental research projects directed to studying genome structure, function, and evolution. Essentially that bioinformatics is recognized as a necessary element for modern experimental research. It is widely used as at the stage of experimental designing and data interpretation, as for solving fundamental problems of organization and evolution of molecular-genetic systems and processes. By analyzing the papers submitted for publication in the two-

volume issues of the BGRS' 2004, the Program Committee came to a conclusion that participants of the Conference have concentrated their attention at consideration of the hottest items in bioinformatics listed below:

computational structural and functional genomics; computational structural and functional proteomics; computational evolutionary biology; computational systemic biology; new approaches to analysis of biomolecular data and processes; bioinformatics and education.

All the questions listed above will be suggested to consideration of participants of BGRS' 2004 at plenary lectures, oral communications, poster sessions, Internet computer demonstrations, and round-table discussions.

BGRS'2004 will host a special "EU-NIS Partnering in Bio-informatics" event, organized by INTAS in close cooperation with the European Commission for activation of international cooperation in the fields of bioinformatics between Russian Federation, other NIS countries, and European Community. The event not only offers chances for meeting the right partner in science or business but also provides the latest information about upcoming calls for proposals in the European Commission's Sixth Framework Programme and the possibilities to jointly apply for these and other grants with colleagues from EU or NIS countries.

Professor Nikolay Kolchanov
Head of Laboratory of Theoretical Genetics
Institute of Cytology and Genetics SB RAS,
Novosibirsk, Russia
Chairman of the Conference

Professor Ralf Hofestaedt
Faculty of Technology
Bioinformatics Department
University of Bielefeld, Germany
Co-Chairman of the Conference

BGRS
2004

COMPUTATIONAL STRUCTURAL AND FUNCTIONAL GENOMICS

SOME WAYS TO INFER A DNA FUNCTION FROM THE SEQUENCE INFORMATION

Abnizova I.*, te Boekhorts R., Gilks W.

MRC-BSU Cambridge, University of Hertfordshire Hatfield, UK
* Corresponding author: irina.abnizova@mrc-bsu.cam.ac.uk

Keywords: *regulatory regions, coding DNA, heterogeneity, computational methods, information entropy, long range correlations*

Summary

We present a computational approach to infer DNA function based on eukaryotic DNA sequence information. Namely, we utilise an observation that exons, regulatory regions and non coding non regulatory DNA exhibit different statistical information patterns. We suggest to capture and measure these patterns with several independent mathematical tools such as rescaled analysis, information entropy, density of low-entropy patches and similarity test.

Results: We introduce here a new optimization technique of which the outcome is independent of the size of the sliding window and hence avoids averaging. This technique, which takes account of the heterogeneity in the DNA sequence, performs an unsupervised search, without using reference sets or cross genome comparison. We also introduce new way of measuring DNA local heterogeneity, and a statistical test for abundance of similar words.
A preliminary application of our methods to the set of genes from six different species, namely human, mouse, fugu, sea urchin, drosophila and yeast, reliably identifies the borders of the regions of interest and thus reveals the potential of our approach.
We propose that this combination of established computational statistical methods, augmented with our sliding window optimization technique, and our new statistical tests, might create a powerful DNA sequence characterization and annotation tool that optimises the search for differences in statistical properties between coding, non-coding and regulatory DNA.
Availability: The software is available from the authors on request.

Introduction

The unsupervised search methods for analysing the structure of the genome fall, broadly speaking, in one of two categories: (i) methods for characterizing the composition of the genome and (ii) those used for detecting sequential or serial dependency (i.e. that focus on the actual ordering of nucleotides).

Nucleotide composition is commonly investigated with tools from information theory (i.e. various ways to estimate the entropy of parts of the genome), self-organising maps (Abe *et al*., 2003), complexity analysis (Wan *et al*., 2003; Li , 1997) and statistical linguistics (Mantegna *et al*., 1994).

Statistical dependencies between nucleotides/amino acids have been analysed using mutual information functions (Azbel, 1995), Markov models (Krogh *et al*., 1994), spectra (Voss, 1992) and methods derived from random walk dynamics such as detrended fluctuation- and rescaled range analysis (Peng *et al*., 1994). In particular the detection of long range correlations (LRCs) has attracted much attention (Li, 1997; Mantegna *et al*., 1994; Azbel, 1995; Peng *et al*., 1994; Herzel *et al*., 1997) and correlations ranging from a few base pairs up to 1000 bp have been found.

Results from the application of these methods seem to indicate that non-coding and coding DNA have distinguishable statistical properties: long range correlations have been reported for non-coding, but not for coding regions (Peng *et al*., 1994; Herzel *et al*., 1997; but see Voss, 1992).

Unfortunately, when these techniques are used in the conventional way, their results will be highly dependent on the size of the sliding window.

Methods and Result

Random walks and rescaled range analysis. One way to characterize the succession of nucleotides is to imagine a "walk" along the DNA string by moving up each time a pyrimidine (a T or C) occurs and by moving down whenever a purine (an A or G) is encountered. As a result, one may obtain a fractal-looking "landscape", in which probably long stretches of mainly purine alternate with stretches that contain mostly pyrimidine. If the probability of the occurrence of a pyrimidine equals that of a purine and is independent of the position in the string (i.e. P (pyrimidine) = P (purine)), then such a "DNA walk" would actually be a "random walk". The *increments* of the walk would form a series of independent and identically distributed events with constant mean and finite variance that is therefore stationary and furthermore characterised by a flat spectrum and the absence of autocorrelations ("white noise"). The random walk itself, being integrated white noise, is typically non-stationary. This is manifest in the largely non-vanishing spikes in the autocorrelation function and the predominance of low frequency components in the power spectrum (Voss, 1992).

Another way to do find long range correlations in DNA sequence is to calculate the Hurst exponent (H) by means of a "rescaled range analysis". The Hurst exponent estimates the degree of "persistence" of the system. Rescaled range analysis can be applied to any time or space series. Setting $x_k = +1$ for k = T, C and $x_k = -1$ for k = A, G, the sequence $\{x_k\}$ can be characterised by:

$$\langle x \rangle_n = \frac{1}{n}\sum_{i=1}^{n} x_i \;,\; X(i,n) = \sum_{m=1}^{i}\left[x_m - \langle x \rangle_n\right]$$

$$R(n) = \max_{i \le n} X(i,n) - \min_{i \le n} X(i,n) \;,\; S(n) = \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \langle x \rangle_n)^2\right]^{1/2}$$

for any $2 \le n \le N$.

For scale free data $R(n)/S(n) \sim (n/2)^H$. Hence, the Hurst exponent (H) can be computed from the least squares fit of the regression of log[R(n)/S(n)] on log[n]. In the above formulas, the test-statistic (H) should be compared with a Hurst exponent obtained under the null-hypothesis that the *cumulative* data (i.e. the time series *after* integration) are from a random walk and therefore that the *original* data are white noise. In other words, $H_{observed}$ should be contrasted with $H_{expected} = 0.5$ and *not* with $H_{expected} = -0.5$ (which would be the Hurst exponent of the original data).

Obviously, changes in persistence are not accounted for by the conventional application of rescaled range analysis, i.e. by using a too large, fixed window. This might be detected as changes in the slope of the log-log plot. We therefore suggest a window that increases until the Hurst exponent reaches a minimum ($H \le 0.5$): this window detects the most probable boards of exons.

Entropy-based DNA segmentation with optimal window method. The conventional procedure for measuring DNA entropy (Li, 1997) typically consists of calculating a frequency vector describing the area's nucleotide composition for a sufficiently large, but subjectively defined area, and then subjecting it to the well-known Shannon function:

$$H_{seq,M} = -\sum_{i=1}^{M} P_i \log(P_i) \,,$$

where M is the length of the frequency vector, which in case of single nucleotides is 4, in case of di-nucleotides is 16, etc. The entropy of a frequency vector is maximal when all elements occur with equal probability, in which case $H_{seq,M} = \log(1/M)$, and hence measures the "eveness" or the

"diversity" of the composition. In contrast, low entropy indicates the "dominance" of a few of the elements. Clearly, when increasingly larger fragments of the same stretch of DNA are used, the entropy of the fragments will asymptotically approach the entropy of the entire stretch. Such an overall estimate does not capture the possibly deviating entropy of small but functionally important subparts. Too large and fixed windows therefore overlook local differences in nucleotide composition.

For a more powerful method, we therefore must optimise the length of the local windows. To this aim, we move sliding window of varying length along the DNA sequence, optimising the length of the window due to the local maximum of Entropy before it reaches asymptotic value. The most high entropy regions are the most likely locations of exons.

Low entropy patches density as heterogeneity measure. There exit two visions of DNA heterogeneity: 1) Heterogeneity is caused by gradual change (bias) of nucleotide composition in different parts of the sequence; 2) Heterogeneity is caused by short runs of repetitive (low-entropy) patterns. If you remove them, DNA will be homogeneous again. It is local heterogeneity.

To calculate the density of low entropy patches, we first developed an entropy-based algorithm able to fish for these patches; this algorithm occurred to detect very weird patterns, which no repeat masker could find. Thus, we used the density of these low entropy patches as a measure of local heterogeneity of DNA.

As the result, we present an example of splitting fugu DNA into functional parts due to low-entropy patches: exons typically have minimum density, non coding non regulatory DNA has maximum density, and regulatory DNA has intermediate density, as one can see at the table below.

In the table, "**diverged"** denotes non coding non conserved DNA: they were picked up randomly through out the Mayfolds fugu whole genome shotgun assembly v.3.0 (August 26, 2002); "**cne"** (conserved non coding elements) are considered as putative regulatory regions: these elements are collected due to ClustalW multiple alignment between fugu, mouse, rat and human in the group of Greg Elgar (Woolfe *et al*., submitted 2004), short elements were concatenated together; exons are randomly picked up in Scaffolds1 and 21 fugu whole genome shotgun assembly v.3.0 (August 26, 2002), www.ensembl.org (Table).

Measure of similar word abundance: similarity-tail test. There is well known observation that there are unusually large (unexpectedly due to random multinomial model) number of some similar words in the regulatory regions. To quantify this fact, we got the collection of known functional regulatory regions from Drosophila genome. For each region, we computed the distribution of clusters of similar words inside it, so called **similarity-tail test.** We calculated the number of similar words of length m with few mismatches exhaustively for each m-word in the given regulatory region. Due to the presence of unusually high number of over-represented words, we expect to have more "fat" (containing a lot of similar words) clusters in comparison with random chance.

To sample this random chance, we shuffled our given sequence 200 times keeping its original di-nucleotide content. For each shuffled sequence we plotted the histograms of similar words distribution. We could easily observe that the original sequence "tails" were significantly longer than all randomised ones.

Almost all regulatory regions analysed (currently, 35) exhibit "fluffy" tails, in contrast with exon regions where there is no significant tail exist. This observation gave raise to the algorithm, which distinguish between exons and regulatory regions due to the presence of statistically significant tails in the original distribution. The algorithm currently is applied for Drosophila genome.

Conclusion: Our collection of tools: **• Information entropy • Rescaled Range Analysis: Hurst exponent• Density of low entropy patterns• Tests for frequency of similar words** might help

to infer the function of a given not yet annotated DNA stretch, and thus to distinguish between different DNA functional parts when all the tools are combined.

In this work we propose an adaptive optimal windowing technique applied for two popular methods which search for short and long range correlations in DNA: rescaled analysis to estimate long distance dependency by the Hurst exponent, and entropy measurement. Both improved techniques are capable to detect putative functional regions in DNA. In addition, we capture DNA homogeneity and presence of repetitive words in other two independent ways: with density of low-entropy patches, and similar words abundance in the DNA stretch. When we combine results of these tests with adaptive window methods, we therefore increase likelihood of detecting functional DNA regions.

Table.

Type of sequence	Density of low-entropy patterns	A + C + G + T	**Pcg** - probability to get **cg**	**Pcg/PcPg** **Pc** - probability to get **c**	Length, bp
Diverged					
Diverged1	0.34	0.29+0.25+0.18+0.26	0.028	0.60	5156
Diverged2	0.44	0.28+0.20+0.21+0.28	0.018	0.41	15221
Diverged3	1.0	0.22+0.36+0.18+0.21	0.03	0.44	755
Diverged4	0.66	0.26+0.18+0.17+0.37	0.02	0.65	1000
Diverged5	0.56	0.30+0.20+0.19+0.30	0.014	0.38	4819
Diverged6	0.62	0.26+0.20+0.20+0.32	0.02	0.48	4750
Diverged7	0.37	0.28+0.20+0.20+0.30	0.02	0.53	8501
Means	**0.58**	**0.28+0.22+0.20+0.29**	**0.02**	0.46	
Cnes					
Cne2	0.032	0.30+0.21+0.18+0.29	0.018	0.44	14782
Cne3	0.0	0.31+0.19+0.19+0.29	0.012	0.34	3423
Cne5	0.018	0.30+0.18+0.19+0.29	0.012	0.33	18876
Cne6	0.078	0.31+0.18+0.20+0.30	0.01	0.29	14500
Cne7	0.02	0.28+0.20+0.21+0.29	0.017	0.39	23258
Cne8	0.05	0.30+0.18+0.21+0.30	0.013	0.34	10378
Means	**0.07**	**0.30+0.19+0.20+0.30**	**0.014**	**0.37**	
Exons					
Ex1_fugu	0.0	0.24+0.27+0.27+0.20	0.048	0.64	4295
Ex2_fugu	0.0	0.27+0.25+0.26+0.22	0.036	0.54	22373
Ex3_fugu	0.03	0.27+0.26+0.26+0.19	0.04	0.57	19940
Ex4_fugu	0	0.25+0.27+0.28+0.20	0.058	0.73	1224
Ex5_fugu	0	0.27+0.26+0.26+0.19	0.045	0.63	3489
Ex6_fugu	0	0.27+0.27+0.26+0.20	0.042	0.58	2835
Ex7_fugu	0.0	0.26+0.27+0.26+0.19	0.043	0.60	3714
Ex8_fugu	0.05	0.27+0.25+0.27+0.20	0.035	0.51	5740
Means	**0.007**	**0.27+0.26+0.27+0.20**	**0.0412**	**0.58**	

References

Abe T., Kanaya S., Kinouchi M, Ichiba Y., Kozuki T., Ikemura T. Informatics for unveiling hidden genome signatures // Genome Res. 2003. V. 13(4). P. 693–702.

Abnizova I., Schilstra M., te Boekhorst R., Nehaniv C.L. A statistical approach to distinguish between different DNA functional parts // WSEAS Transactions on Computational Methods. 2003. V. 2. Issue 4. P. 1188–1196.

Azbel Y.M. Universality in a DNA statistical structure // Physical Review Letters. 1995. V. 75. P. 68–171.

Herzel H., GroЯe I. Correlations in DNA sequences: the role of protein coding segments // Physical Review E. 1997. V. 55. P. 800–810.

Krogh A., Mian S., Haussler D. A hidden markov model that finds genes in *E. coli* DNA // Nucleic Acids Res. 1994. V. 22. P. 4768–4778.

Li W. The complexity of DNA // Complexity. 1997. V. 3. P. 33–37.

Mantegna R.N., Buldyrev S.V., Goldberger A.L., Havlin S., Peng C.K., Simons M., Stanley H.E. Linguistic features of noncoding DNA sequences // Physical Review Letters. 1994. V. 73. P. 3169–3172.

Peng C.K., Buldyrev S.V., Havlin S., Simons M., Stanley H.E., Goldberger A. Mosaic Organization of Nucleotides // Physical Rev. E. 1994. V. 1. P. 1685–1689.

Voss R. Evolution of Long-Range Fractal Correlations and 1/f Noise in DNA Base Sequences // Physical Review Letters. 1992. V. 68. P. 3805–3808.

Woofle A., Goodson M., Goode D., Snell P., Smith S., Vavouri T., McEwen G., Gilks W., Walter K., Edwards Y., Elgar G. Highly conserved non coding sequences are associated with developmental control genes in vertebrates, submitted 2004.

Wan H., Li L., Federhen S., Wootton J.C. Discovering simple regions in biological sequences associated with scoring schemes // J. Comput Biol. 2003. V. 10(2). P. 171–85.

AUTOMATIC LANE DETECTION AND SEPARATION IN ONE DIMENSIONAL DNA GEL IMAGES

Akbari A.**, *Algregtsen A.

Department of Informatics, University of Oslo, P.O.Box 1080, N-0316 Oslo, Norway
* Corresponding author: e-mail: akbara@ifi.uio.no; fritz@ifi.uio.no

Keywords: *lane letection, lane separation, DNA*

Summary

Motivation: In this paper we describe the automatic detection and separation of lanes in DNA gel images, applying an iterative low-pass filter and equivalent width algorithm. DNA gel images often contains more than one lane in which, each lane contains either DNA for an individual or standard markers. Each time DNA for an individual in one lane is going to be analyzed or compared with DNA from other individuals within or between gel images, the location of each lane within the gel images should be recomputed. Isolating the lanes in a DNA gel image and making a specific image file for each lane makes the process of analyzing and the comparison of individuals more feasible. In addition the storage space will also be reduced.

Results: A simple iterative low-pass filter followed by equivalent width algorithm is applied to localize and separate the lanes in DNA gel images. Results obtained by this method are reliable and will reduce further computations when the DNA for an individual is going to be compared with other individuals within or between gel images.

Availability: available on request from the authors.

Introduction

DNA gel images often contain more than one lane, where each lane contains DNA for one or more (mixed samples) individuals. The efficiency of both within and between gel comparisons will be increased by having DNA fingerprints for each individual on a separate image (file). That is because the data for each lane (individual) would be available directly from its image file instead of searching through the whole gel image every time it is needed.

The focus of this work will be applying an iterative low-pass filter and equivalent width algorithm to locate and separate lanes in one dimensional gel images. A number of pre-processing steps should be applied on the image prior to the lane detection and separation. First a low-pass homomorphic filter is applied to enhance the image followed by an edge preserved noise filtering algorithm. Then a background normalization operation is applied to the resulting image. Gel electrophoresis and different probes for hybridization are used to produce the one dimensional DNA fingerprints and the images are generated on X-ray film (Kirby, 1990; Burke, 1989; Debenham, 1992). Ideally, the lanes on the electrophoresis should be equidistant. But, DNA are often degraded by different types of noise, e.g. film grain noise and noise from other sources, such as electrophoresis, membrane, the hybridization process, etc. Due to these degrading facts, lanes on the gel images are not equidistant (Fig. 1), and therefore the process of separating lanes from gel images will be complicated. The gel images we are working with have no non-migrated bands that represent the electrophoresis wells which could be used as a reference point for each lane. On the contrary we have no a-priori information for the start and the end of each lane. Sometimes, handling the gel into the membrane will also cause localized stretching on the gel. If such kinds of distortions are too large, the image should be adjusted using a suitable geometrical transformation algorithm. Otherwise, dropping a few pixels (depending on the scanning resolution) from each side of each lane will help to separate the lanes from the gel image. The gel image in Fig. 1a indicates that there are no specified edges in the vertical direction for lanes in the image.

The width (X-direction) of the DNA bands on each lane varies due to their radioactive strength. The spaces between lanes are almost indistinguishable on the lower part of the image. Since the lanes on the DNA fingerprint gel images have no specified edges, artificial edges will be made for each lane on the image. To do so, a 2-D. mean filter of the size (3, ImageHeight/3) is used to smooth the DNA bands along each lane (Fig. 1b).

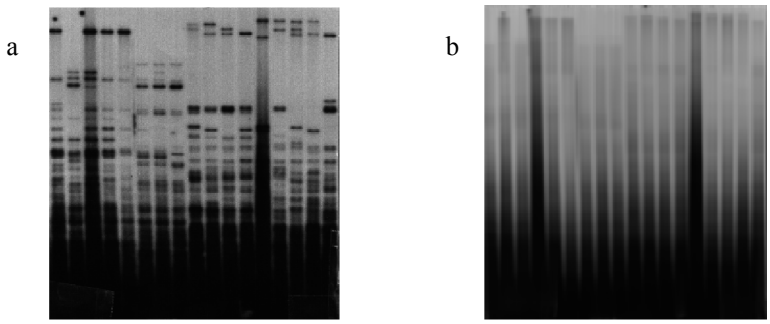


Fig. 1. (a) A DNA fingerprint image contains lanes which are not equidistant due to noise and other effect in the image; (b) the image obtained by using a two dimensional mean filter of the size (*3, ImageHeight/3)* on image (a).

Methods and Algorithms

Iterative Low-pass Filter. An approach for separating the lanes from a gel image is based on the one dimensional signal obtained by averaging the intensity for each column in the gel image into the horizontal axis. Figure 2, illustrates the average signal computed from the gel image in Fig. 1a.

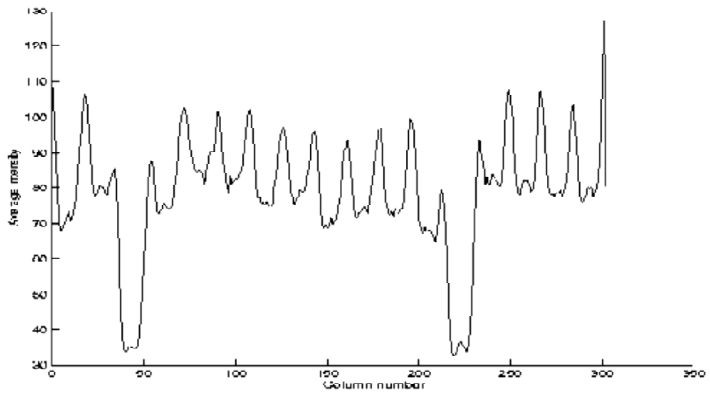


Fig. 2. Signal obtained by averaging the intensity for each column in the image in Fig. 1b.

To be able to use the average signal for computing the positions of the lanes in the gel image, the average signal should be smoothed such that the number of minimum points on the signal being equal to the number of lanes in the image. At the same time the minimum points in the smoothed signal should be approximately equidistant. The number of the lanes in the image is given as input parameter to the program. The approximate distance between two lanes in an image with a sampling frequency of one pixel per 200 μm will be 7–8 pixels. This option can also be changed by the user. Total space between lanes in the image will be approximated by multiplying the (number of (lanes – 1)) with the approximate distance between lanes. The lane width which is

used to compute the window size for the first iteration of low-pass filter is approximated with dividing the effective width by the number of lanes.
The effective width is: (*image width*) – (*total distance between lanes*).
To smooth the average signal a low-pass filter with a window size ((*4/5*)*(*Lane width*)) is used. The window size is chosen to be smaller than the approximate lane width to preserve the information for the lanes in the gel image. We use this argument to generalize this filter applicable on different DNA fingerprint images. Otherwise different window sizes should be chosen for different images. The low-pass filter is used because the low frequency part of the average signal assumed to represent the lane frequencies in the image and the higher frequencies in the signal are assumed to noise. The smoothed signal will be analyzed, if the number of minimum points in the resulting signal is more than the number of lanes on the gel image, the smoothing operation will be performed iteratively until the correct result is obtained. If iterative smoothing is acquired, then the size of the window for next iteration will be adjusted with the factor of (*1/3*)*(*previous window size*). The reason of choosing window sizes in this way is to preserve the information about the lanes in the image.
This could be written as follow:
*W(0)= (4/5)*LaneWidth*
*W(n)= (1/3)*W(n-1)* where, n = 1, ...
The procedure for computing the approximate lane width the following algorithm is used.
Total distance between lanes = (Number of lanes - 1) * (Approximate distance between lanes).
Approximate effective width = (Image width) - (Total distance between lanes).
Approximate lane width = (Approximate effective width) / (Number of lanes).
As it is outlined above, the resulting lane width will be used to compute the size of the window for the low-pass filter in the first pass. The signal obtained by this filtering process will be analyzed to find out if more filtering iterations should be performed. Each local minimum of the smoothed signal in Fig. 3, a represents a point within the corresponding lane in the image in Fig. 1a. To visualize this point, it is illustrated by "white line" along each lane in the image in Fig. 3b.

Information, such as coordinates and intensity levels for each local minimum and its two neighboring maximum points are used to compute the edge coordinates of the corresponding lane in the image. Computing of edge coordinates for each lane is performed by using the equivalent width algorithm (Gray, 1976; Oppenheim, 1983; Marple, 1987; Pristley, 1989). The results of applying the iterative low-pass filtering followed by the equivalent width algorithm on three different DNA fingerprint images are illustrated in Fig. 4.

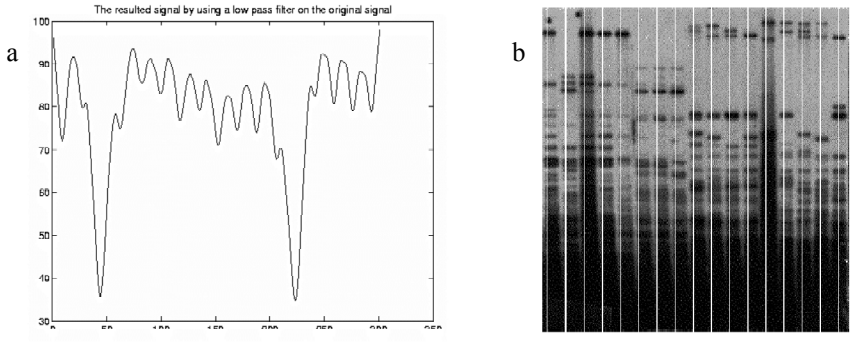


Fig. 3. (a) Signal obtained by using the low-pass filter with window size W(0)=9 (computed by the procedure described in the text) on the signal in Fig. 2 in only one iteration; (b) *"White line"* along the lanes in the image represent the corresponding local minima of the smoothed signal in (a).

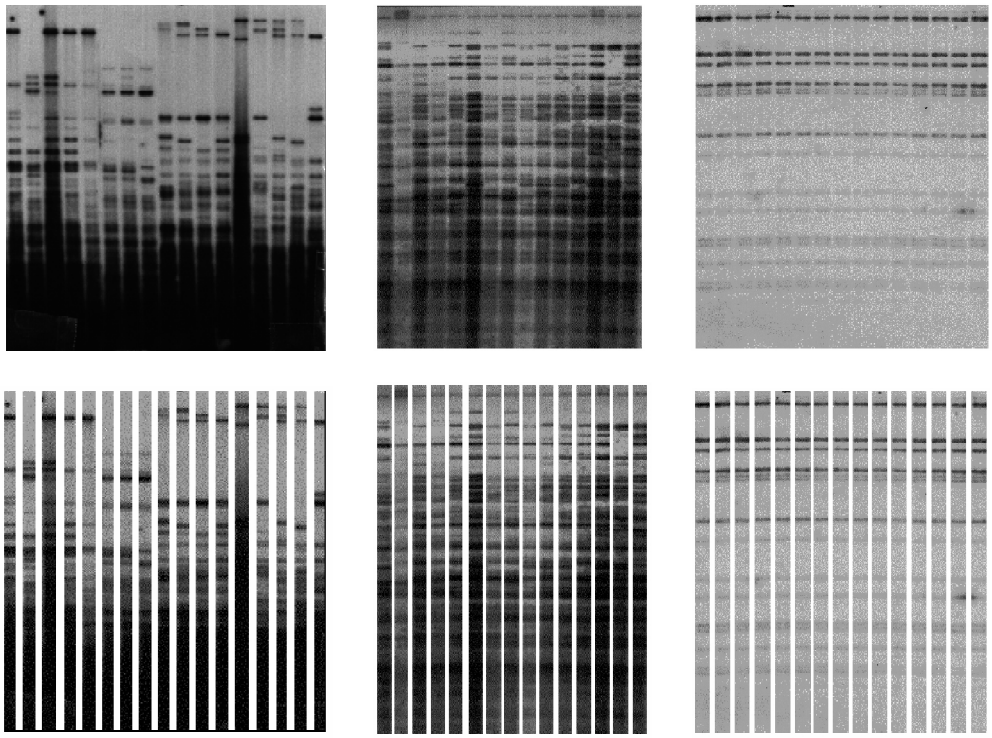


Fig. 4. Images on the second row are obtained by applying the iterative low-pass filtering followed by the equivalent width algorithm on the images on the first row correspondingly.

Discussion

In this work an iterative low-pass followed by equivalent width algorithm is applied to detect and separate lanes in gel images. Results obtained by this method are reliable, but it has some disadvantages. One disadvantage to this method is the approximate distance between two lanes which is the user choice and makes effect on the separated number of lanes. If the number of lanes resulted from the program compared to the number of lanes (in the image) input to the program are different, then the separated lanes should be analyzed visually about the lane/lanes which are not detected.

References

Burke T. DNA Fingerprinting and Other Methods for the Study of Mating Success. Tree, May 1989. V. 4, N 5. P. 139–144.

Debenham P.G. Probing Identity: The changing face of DNA fingerprinting. Tibtech march 1992. V. 10. P. 96–102.

Gray D.F. The observation and analysis of stellar photospheres. John Willy & Sons Inc. 1976. P. 259–334.

Kirby L.T. DNA fingerprinting. An introduction. Stockton press. 1990.

Marple L.S. Digital Spectral Analysis with Applications. Prentice-Hall Signal processing Series. 1987.

Oppenheim A.V., Willsky A.S., Young I.T. Signals and Systems. Prentice Hall International Inc. 1983. P. 277–278.

Pristley M.B. Spectral Analysis and Time Series. Probability and Mathematical Statistics. Academic Press. London England. 1989. P. 517–528.

GpiMap, AN ENVIRONMENT FOR GENETIC/PHYSICAL MAP DATA MANAGEMENT, VISUALIZATION AND COMPARATIVE ANALYSIS

Albini G.*[1], *Chetouani F.*[1]*, *Rouille S.*[1], *Karsenty E.*[1], *Thomas B.*[2], *Legeai F.*[1], *Samson D.*[1], *Pereira L.*[5], *Arcade A.*[3], *Joets J.*[3], *Scala D.*[2], *Viara E.*[5], *Barillot E.*[4], *Duclert A.*[1]

[1] GenoplanteInfo, Unite de Recherche Genomique-Info, INRA, 523, place des Terrasses de l'Agora, 91000 EVRY, France; [2] RhoBioInf, 2 rue Gaston Cremieux, C.P. 5707, 91057 Evry Cedex, France; [3] INRA, Ferme du Moulon, 91190 Gif-sur-Yvette, France; [4] Institut Curie, 26 rue d'Ulm, 75248 Paris Cedex 05, France; [5] SYSRA, 523, place des Terrasses de l'Agora, 91000 EVRY, France

* Corresponding authors: e-mail: gpinterface@infobiogen.fr

Keywords: *comparative genomic, genetic mapping, marker, QTL, database*

Summary

Motivation: For plant studies, comparative genomics is applied to transfer structural and functional knowledge from model genomes to related genomes.

Results: To support the biologist in this comparative genomic task, we have set up an environment for genetic/physical map data management, visualization and comparison: GpiMap.

Availability: GpiMap access will be soon available at the URL http://genoplante-info.infobiogen.fr

Introduction

Identification of genes involved in quantitative traits is a first step to decipher biological and physiological mechanisms underlying traits. According to comparative studies, in plant species families in spite of large genome size variation, gene organization is more conserved than anticipated (Gale, Devos, 1998; Keller, Feuillet, 2000). So, for plant genomes not currently available, an important strategy is to use information inferred from complete or nearly complete related plant genome models taking advantage of known genome colinearity (synteny).

In this context, we have developed an environment for genetic/physical map data management, visualization and comparison: GpiMap.

Implementation and Results

The GpiMap environment currently includes a mapping database, a textual web interface for the database and a map comparison tool.

Mapping database. A mapping database, stored in the Oracle RDBMS has been set up to store all types of mapping information from different species: genetic maps of markers, Restriction fragment length polymorphism (RFLP), microsatellites, deletion bin maps, Quantitative Trait Loci (QTL), maps, populations and all information linked (Experimental conditions, QTL phenotypic evaluation, ...).

The database currently contains maize, wheat and rice genetic maps, as well as QTLs for some species. New extension of the database schema will concern synteny data and physical maps.

Textual web interface. A web-based text query interface allows extended queries in the mapping database. Query web forms include the possibility to query the database by loci, QTLs, markers, maps and traits. The tabulated resulting output provides links in both directions between markers and loci, and most importantly, between QTLs and loci, allowing the biologist to easily study the context of its favourite QTLs.

Last developments include graphical query outputs: for instance a loci tabulated output can be represented as a set of map drawings. A Blast search form is available to query the database of all the sequences anchored on genetic maps.

The whole web application described is based on JSP/Servlet technologies using the model-view-controller framework Struts. Database querying layer relies on the free Object/Relational Mapping tool Hibernate.

Map comparison tool. A web interface called MapComparator provides the ability to graphically compare genetic maps. It proposes to select maps either from the database or from local uploaded files. The user can select a subset of linkage groups and a subset of biological objects mapped on them. After this selection process, the tailored maps are transferred to a Java applet for graphical display and comparative purposes.

The applet offers a dynamic view of several maps side by side to allow their easy comparison. It displays in the same view both QTLs and loci so that their positions can be directly compared. The maps order can be changed, common markers can be highlighted and connected by lines for synteny study purposes, zoom level can be adjusted independently for each map representation and the number of displayed markers adapts to the zoom level. To get information about a locus or QTL, a mouse click on its representation opens the associated textual web site page in a browser window.

Last developments include export file facilities. To study comfortably syntenic regions, the new link/unlink tool allows the scrolling and zoom features to operate on the maps together (linked) or separately (unlinked).

Enhancements will include deletion map and physical map support, and edition features.

This application relies on an RMI (Remote Method Invocation) architecture for the data transfert between our server and the user's applet.

Discussion

These integrated tools should greatly accelerate the discovery of interesting genes by allowing seamless navigation between genetic maps and physical maps, as well as easy comparison between model species and economically important species.

Acknowledgements

This work is supported by the french Genoplante program (http://www.genoplante.com). We are very grateful to the Genoplante mapping workgroup. In particular we would like to express our gratitude towards Fabien Chardon, Matthieu Falque, Marie-Henriette Flament, Philippe Leroy, Jean-Pierre Martinant.

References

Gale M.D., Devos K.M. Plant comparative genetics after 10 years // Science. 1998. V. 282. P. 656–569.

Keller B., Feuillet C. Colinearity and gene density in grass genomes // Trends Plant Sci. 2000. V. 5. P. 246–251.

DATABASE OF LONG TERMINAL REPEATS IN HUMAN GENOME: STRUCTURE AND SYNCHRONIZATION WITH MAIN GENOME ARCHIVES

***Alexeevski A.V.*[1], *Lukina E.N.*[1], *Salnikov A.N.**[2], *Spirin S.A.*[1]**

[1] Belozersky Institute, Moscow State University, Moscow, Russia; [2] Department of Computational Mathematics and Cybernetic, Moscow State University, Moscow, Russia
* Corresponding author: e-mail: salnikov@angel.cs.msu.su

Keywords: *human genome, long terminal repeats, synchronization, data bases*

Summary

Motivation: The complexity of main nucleotide and complete genome databases (DBs) makes it essential to develop special well-structured, equipped by user-friendly interface, and curated by experts secondary DBs. The problem of synchronization of the information between DBs arises. This problem seems to be especially actual for the human genome because of its continued improving and ordering.

Results: Yu.B. Lebedev group, Shemyakin and Ovchinnikov Institute of Bioorganic Chemistry, collected and annotated long terminal repeats of class 5 (LTR5). This collection was the starting point for the development of Human Endogenous Repeat database (HEREbase). Currently, HEREbase contains nearly all (~1000) appearances of LTR5 objects in human genome. Query forms allows user to select specified entries. The created program scripts weekly synchronize links to relevant objects in the nucleotide DB (EMBL) and human genome DB (EnsEMBL).

Availability: http://math.genebee.msu.ru

Introduction

The sequences of the complete human genome are contained in well-known publicly opened data bases like the archive database of nucleic acid sequences EMBL (http://www.embl.org) and the database of complete genomes EnsEMBL (http://www.ensembl.org). Such large universal archives are rather complicated and not convenient for purposes of studying specific objects in genome. A user can have difficulties to create adequate queries; searching for results, which can contain thousands of sequences, takes a lot of time due to DB size and multiple simultaneous users requests. In addition, an expert hardly can add new information about an existing object to an archive DB. These inconveniences can be reduced by creating smaller specialized secondary databases that are well-structured and equipped by user-friendly interface. Due to the continued reordering of the human genome the problem of synchronization of the information in small specialized DB with the information in archive DBs becomes very important and somewhat nontrivial.

One of intensively investigated objects in human genome are so-called Long Terminal Repeats (LTR), which are traces of ancient retroviral invasions. The potential possibility of LTRs to be involved into gene expression regulation makes them interesting for scientists who study human evolution, inherited diseases, cancerogenesis etc. (Khodosevich *et al*., 2002). A curated DB for LTRs and neighboring genes can be useful for such studies.

Methods and Algorithms

HEREbase was created in *MySQL* under Linux (debian) operating system. The user and curator interfaces are made by means of *Apache* and *php4*. Scripts to synchronize the information in HEREbase with the databases EMBL and EnsEMBL, are written in Python language.

Results

We developed a small specialized DB named Human Endogeniouse RetroElement database (HEREbase, http://math.genebee.msu.ru). Currently, the HEREbase is filled by one class of LTRs, namely LTR5. An annotated collection of LTR5 created by E.D. Sverdlov and Yu.B. Lebedev (Shemyakin – Ovchinnikov Institute of Bioorganic chemistry), was in the background of HEREbase. It was extended to all annotated LTR5 objects in EnsEMBL, using information from HERVd (http://herv.img.cas.cz/, see Pačes, Pavliček *et al*., 2004). The HEREbase is filled also by neighboring genes of LTRs.

HEREbase relational tables contain information on LTRs themselves: their sequence, location in the genome (chromosome, arm, band), 5' and 3' short direct repeat sequences, class of the LTR (according to the classification of Lavrentieva *et al*., 1998), and links to the relevant EMBL entry and to the coordinates in the chromosome according to the latest release of EnsEMBL. Additionally, HEREbase contains flanking sequences to facilitate the identification of the LTR, the information about genes in 100k neighborhood and repeats in 2k neighborhood. HEREbase can be curated via an Internet browser. For this purpose, a user ('guest') and a curator have different privileges. HEREbase queries suggests a number of possibilities for LTR5 selection. It is possible to make restrictions on classes of LTR5, on involvement/non-involvement of a LTR5 in a provirus, on chromosome localization (e.g., "8p12"). Also LTRs can be selected that are close to annotated genes. User can input a segment of LTR sequence or a sequence of an LTR's flank and possess the list of LTR5 that have this segment in their sequences (resp. sequences of their flanks). If user does not input any restrictions, then he gets the list of all LTRs in the data base. To synchronize the information, a number of program scripts in Python language were created. Each script weekly automatically checks specific links to EnsEMBL or EMBL and extracts the information specified in HEREbase. In the case of disagreement a script tries to correct discordance. If it is possible, the corrected information is put into HEREbase. If no, the script includes the information about the disagreement into the special table visible (via http) to curators of HEREbase. A curator of the HEREbase is supposed to resolve the found conflict. In the case of EMBL the links the script finds are an EMBL entry by accession number specified in HEREbase and coordinates of the LTR in the entry, which are corrected (if needed) using BLAST. More complicated algorithm is needed to deal with EnsEMBL. EnsEMBL is a *mysql* operated relational database whose tables are open via *ftp* protocol. The table cross-references reflects complicated relations between all detectable objects in the genome. For example, LTR coordinates in a chromosome can be derived only from the third level of EnsEMBL table inclusions. An additional problem arises after correction of EnsEMBL structure by the EnsEMBL programmer group. For example, the table named 'contigs' in release 19 was transferred into the table named 'sequence_region' in the release 20. Such EnsEMBL corrections requires corrections in the HEREbase scripts. Currently HEREbase contains 1032 LTR5 entries. The process of annotation of these entries is continued. The HEREbase structure and software can be used not only for LTR5 in human genome, but also for other classes of repeats in any complete eukariotic genomes.

Acknowledgements

The authors are grateful to Yu.B. Lebedev and I. Mamedov (Shemyakin – Ovchinnikov Institute for Bioorganic Chemistry RAS) for the background data and useful consultations. The work was supported by Ludwig Institute for Cancer Research (CRDF grant RB01277-MO-2).

References

Khodosevich K., Lebedev Y., Sverdlov E. Endogenous retroviruses and human evolution // Comparative and Functional Genomics. 2002. V. 3. P. 494–498.

Lavrientieva L., Khil P., Vinogradova T., Akhmedov A., Lapuk A., Shakhova O., Lebedev Y., Monastyrskaya G., Sverdlov E.D. Subfamilies and nearest-neighbour dendrogram for the LTRs of human endogenous retroviruses HERV-K mapped on human chromosome 19: physical neighborhood does not correlate with identity level // Human Genetics. 1998. V. 102. P. 107–116.

Pačes J., Pavliček A., Zika R., Kapitonov V.V., Jurka J., Pačes V. HERVd: the Human Endogenous REtroViruses Database: update // Nucleic Acids Res. 2004. 32:D50.

RECOGNITION OF CODING REGIONS IN GENOME ALIGNMENT

***Astakhova T.V.*[1], *Petrova S.V.*[1], *Tsitovich I.I.*[2], *Roytberg M.A.**[1]**

[1] Institute of Mathematical Problems in Biology RAS, Puschino, Russia; [2] Institute of Information Transmission Problems RAS, Moscow, Russia

* Corresponding authors: e-mail:, Roytberg@impb.psn.ru

Keywords: *coding region, gene recognition, genome alignment, synonymous and non-synonymous substitution*

Summary

Motivation: Gene recognition is an old and important problem. Statistical and homology based methods work quite well, if one attempts to find long exons or full genes, but is unable to recognize relatively short coding fragments. Genome alignments and study of synonymous and non-synonymous substitutions provide opportunity to overcome this drawback. Our aim is to propose a criterion to distinguish short coding and non-coding fragments of genome alignment and to create an algorithm to locate alignment coding regions.

Results: We developed a method to locate aligned exons in a given alignment. First, we scan the alignment with a window of a fixed size (~40 bp) and assign the score H(P) to each window position P a score H(P). The value H(P) reflects, if numbers K_S of synonymous substitutions, K_N of non-synonymous substitutions and D of deleted symbols appear to be similar for coding regions.

Second, we mark "exon-like" regions, ELRs, i.e. sequences of consecutive high-scoring windows. Presumably, each ELR contains one exon. Third, we highlight an exon within every ELR. All the steps have to be performed twice, for the direct and reverse complementary chains independently. Finally, we compare predictions for two chains to exclude possible predictions of "exon shadows" on complementary chain instead of real exons. Tests have shown that 93 % of marked ELRs have intersections with real exons and 93 % of aligned annotated exons intersect marked ELRs.The total length of marked ELRs is ~ 1.35 of the total length of annotated exons. About 75 % of the total length of predicted exons belong to annotated exons and more, than 70 % of the total length of correctly aligned annotated exons belong to predicted exons. The run-time of the algorithm is proportional to the length of a genome alignment.

Introduction

The existence of powerful genome alignment methods (Brudno *et al.*, 2003; Roytberg *et al.*, 2002) and availability of many complete genomes, including several eukaryotic, lead to revisions of classical problems of sequence analysis. Indeed, we can analyze pair-wise (or, if possible, multiple) sequence alignment instead of one genome sequence. In the case of the gene recognition problem, genome alignments allow to take advantage of two ideas. First, coding regions are, in general, more conserved than non-coding. Thus, one can attempt to recognize genes as a sequence of well-aligned genome fragments (Bafna, Huson, 2000; Batzoglou *et al.*, 2000; Novichkov *et al.*, 2001; Taher *et al.*, 2003). Such methods are effective for relatively distant species, but some genes can be unrecognizable because of low similarity between species. On the other hand, alignment of close genomes often gives many false positive exons, because of the existence of conserved non-coding regions (Shabalina, Kondrashov, 1999). Second, one can additionally pay attention to the difference between substitution patterns in the coding and non-coding regions, the former tend to be synonymous, i.e. to preserve a coded residue. The metods using alignment-based HMMs or pair HMMs (Meyer, Durbin, 2002; Pedersen, Hein, 2003), take into account the differences between various parts of a genome alignment implicitly, in course of HMM training. Despite the promising results, shown by the methods, we think that it is worth to learn explicitly,

what benefit one can get from the differences in substitution patterns. The explicit usage of the differences is implemented in (Nekrutenko *et al.*, 2001), abilities of the approach were demonstrated in (Nekrutenko *et al.*, 2002). However, the goal of the paper (Nekrutenko *et al.*, 2001) is mainly to study the ability of the proposed criterion to recognize relatively long exons as a whole, authors make no attempt to recognize exon borders or short coding regions.

We propose a two-stage procedure combining prediction techniques of traditional identification of exons in DNA sequence and methods based on information about genome alignment. First, using investigation of substitution patterns, we perform an alignment filtration, i.e. locate "exon-like regions" (ELR) in the alignment. Then, the putative exon within ELR can be found using the classical statistical approach. Below we will demonstrate the advantages and drawbacks of the approach and will discuss possible ways for improving it.

Materials and Methods

General description of the approach. The algorithm works in four steps. The first three have to be performed independently for direct and reverse complementary complement chains. At the last step we compare the results obtained for two chains and prepare the final prediction. We start (the first step) with scanning the alignment with a window of a fixed size w and a given shift s. For each considered window, we make the decision if it is exon-like or not. Then (the second step), we reveal "exon-like" regions, ELRs. An ELR is a set of consecutive window positions (see details below). Any two ELRs marked on a chain do not intersect each other. Presumably, each ELR contains one exon. At this step, we work only with exon/nonexon marks of window positions, the marks were assigned on previous step. On the third step we reveal a putative exon for each ELR and assign a score to the exon. If ELR does not contain a pair assigned a aligned exons of high enough score, the ELR is to be rejected. Finally, we compare ELRs found on the direct and reverse chains. If two ELRs of different chains intersect, we keep only one, the ELR having an exon assigned a higher score.

Analysis of window position. Let w be the size of the window. For a window at the position P, i.e. for the fragment of alignment from position P to position P+w-1, the program calculates its score H(P). The score characterizes the presence of stabilizing selection at the protein level. The basic characteristics of the window at a given position of alignment are: 1) FMatch – fraction of match alignment positions, i.e. superposition of identical nucleotides; 2) Probability $Pr(K_T, K_S, D)$ to obtain K_S or more synonymous substitutions, if K_T random independent substitutions were performed and D codons are deleted. We calculate $Pr(K_T, K_S, D)$ for three possible frames. The score H(P) is a negative binary logarithm of the minimum of the three probabilities. The window position P is "exonic", if both FMatch(P) and the score H(P) exceed a threshold.

Exon-like regions. A region is a set of consecutive windows, i.e. windows at positions P, P+s, P=2s, …, where s is a given shift. A region starts at the beginning of the 1st window and ends at the end of the last window. An exon-like region (ELR) is a region meeting the following conditions: 1) The 1st window of the region contains a putative acceptor site or START codon; the last window of the region contains a putative donor site or STOP codon (see below); 2) A region is "exone-dense", i.e. the difference between the number of non-exonic and exonic windows within a consecutive part of a region cannot exceed a threshold InnerCut; 3) The number of non-exonic windows at the beginning and at the end of a region cannot exceed a threshold EdgeCut; 4) a region is not a part of another fragment meeting conditions 1– 3.

A putative acceptor (donor) site is aligned, i.e. present in both sequences, dinucleotide "AC" ("GT"), having Berg-von Hippel score (Berg, von Hippel, 1987) exceeding given cut-off. Putative START- and STOP-codons also have to be presented in both sequences and to be aligned.

Putative exons. Putative exon is a part of an exon-like region, starting with an acceptor or START

and ending with a donor or STOP. We assign a statistical score S(E) and alignment score A(E) to every putative exon E. The score S(E) is the sum of scores of exons, given on every genomic sequence, calculated by the method described in (Gelfand *et al.*, 1996). The value S(E) depends on scores of splicing sites, codon potential and exon length. Alignment score A(E) reflects the difference between the ratios K_S/K_N for the exon calculated for the considered chain and the reverse chain. The score G(R) of an ELR is the sum of the maximum values of S(E) and A(E) for putative exons belonging to the region. If the value G(R) is below a cut-off (currently, 2.25), the region R will be rejected. Otherwise, the exon corresponding to the maximum score S(E) is considered as a predicted exon for the region R.

Genome alignment and gene annotation. The program implementation of the method was tested on the alignment of syntenic regions of the 6th *Homo sapiens* chromosome and the 17th *Mus musculus* chromosome of ~700000 nucleotides length. The alignment was obtained by the OWEN program (Ogurtsov *et al.*, 2002). There are 56 genes annotated on the human sequence and 58 genes annotated on the mouse sequence. Alternative splicing variants are given for 17 human genes, and for only one mouse gene. Mouse genes contain 567 annotated exons, 479 of these are aligned correctly with the corresponding human exons. Incorrect alignment of the other genes can be mostly explained by the inconsistency of exon the annotation in the human and mouse genome. The total length of all the annotated mouse exons is 100,869, the average exon length is 178.

Testing parameters. We have used the following parameters values (see above): 1)Window size w = 40; window offset s = 10 bp; 2) FMatch cut-off for "exonic" window FMatchMin = 0.65; H(P) cut-off for "exonic" window H_Min = 2.25; 3) ELR cut-offs InnerCut = 6; EdgeCut = 6; 4) minimum score of an acceptor splicing sites ACC_Score = -17; minimum score of a donor splicing site DON_Score $\geq$ -7; 5) the minimum length of an inner exon is 40 bp; the minimum length of start or exon is 15 bp.

Results and Discussion

Testing results. The algorithm produces two types of objects (see Materials and Methods): exon-like regions (ELR), and putative exons. Below we give all the results related to the mouse chromosome. The results for the human chromosome are very similar.

We have revealed 628 ELRs, the total length of a revealed ELR is 13,694 (136% of the total length of annotated exons), the average ELR length is 218 (122 % of the average exon length). 586 of these (93 %) have intersection with an annotated exon. Only 35 of the 479 correctly aligned exons (7 %) do not intersect ELRs.

Putative exons, highlighted within ELRs, show the following results. 408 correctly alignment annotated exons (or 85 %) intersect corresponding putative exons. Usually, the intersection between putative and annotated exon cover almost all the annotated exon (87 % on average), 215 correctly aligned exons (50 %) are recognized correctly. At least one exon border is recognized correctly for 406 exons (84 %). The common part of all the correctly aligned exons and corresponding putative exons constitute 74 % of the total length of correctly aligned annotated exons. Approximately the same part of the total length of putative exons belongs to annotated exons.

Discussion. The algorithm addresses two problems. First, it approximately locats the area cohere it is reasonable to search for exons (generation of exon-like regions). Second, highlights out putative exons within the ELRs. The problems are relatively independent, i.e. we can use an arbitrary gene recognition algorithm to solve the second problem, when the first is already solved.

Our main efforts were targeted at the first problem, and the algorithm efficiently solve it. Taking into account its linear run-time, the algorithm can serve as a useful filtration tool for any exon-recognition algorithm, working with genome alignments. We envigage ways of improving the method. In particular, we plan to gather statistics of the possible values of pairs (K_N, K_S) in the

coding and non-coding regions, and define the scoring function H(P) using the maximum likelihood principle.

Putative exons show a weaker correlation with annotated exons than the ELRs. We plan to improve significantly this part of the algorithm. For example, we plan to generate for a given ELR several putative exons having different frames and the link them to predict the whole gene. Another possible development of the project is to realign genomes in the vicinity of putative exon borders. General genome alignment algorithms often misalign conservated positions of splicing sites.

Aknowledgements

Authors are thankful to A.Kondrashov, A.Ogurtsov, S.Shabalina and S.Sunyaev for helpful discussions. The work has been supported by the Russian Found for Basic Research (project nos. 03-04-49469, 02-07-90412) and by grants from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003) and NWO.

References

Bafna V., Huson D.H. The conserved exon method for gene finding // Proc. Int. Conf. Intell. Syst. Mol. Biol. 2000. V. 8. P. 3–12.

Batzoglou S., Pachter L., Mesirov J.P., Berger B., Lander E.S. Human and mouse gene structure: comparative analysis and application to exon prediction // Genome Res. 2000. V. 10(7). P. 950–8.

Berg O.G., von Hippel P.H. Selection of DNA binding sites by regulatory proteins.Statistical-mechanical theory and application to operators and promoters // J. Mol. Biol. 1987. V. 193. P. 723–750.

Brudno M., Malde S., Poliakov A., Chuong B. Do, Couronne O., Dubchak I., Batzoglou S. Global alignment: finding rearrangements during alignment // Bioinformatics. 2003. V. 19. P. 54–62.

Gelfand M.S., Podolsky L.I., Astakhova T.V., Roytberg M.A. Recognition of genes in human DNA sequences // J. Mol. Biol. 1996. V. 3, N 2. P. 223–234.

Meyer I.M., Durbin R. Comparative ab initio prediction of gene structures using pair HMMs // Bioinformatics. 2002. V. 18(11). P. 1546–1547.

Nekrutenko A., Chung Wen-Yu., Li Wen-H. An evolutionary approach reveals a high protein-coding capacity of the human genome // Trends in Genetics. 2003. V. 19, N 6. P. 306–310.

Nekrutenko A., Makova K., Wen-Hsiung Li. The K_A/K_S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study // Genome Res. 2001. V. 12. P. 198–202.

Novichkov P.S., Gelfand M.S., Mironov A.A. Gene recognition in eukaryotic DNA by comparison of genomic sequences // Bioinformatics. 2001. V. 17(11). P. 1011–8.

Ogurtsov A.Y., Roytberg M.A., Shabalina S.A, Kondrashov A.S. OWEN: aligning long collinear regions of genomes // Bioinformatics. 2002. V. 18. P. 1703–1704.

Pedersen J.S., Hein J. Gene finding with a hidden Markov model of genome structure and evolution // Bioinformatics. 2003.V. 19(2). P. 219–27.

Roytberg M.A., Ogurtsov A.Y., Shabalina S.A., Kondrashov A.S. A hierarchical approach to aligning collinear regions of genomes // Bioinformatics. 2002. V. 18. P. 1673–1680.

Shabalina S.A., Kondrashov A.S. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes // Genet. Res. 1999. V. 74. P. 13–22.

Taher L., Rinner O., Garg S., Sczyrba A., Brudno M., Batzoglou S., Morgenstern B.A. AGenDA: homology-based gene prediction // Bioinformatics. 2003. V. 19(12). P. 1575–7.

ALGORITHM FOR SEARCHING FOR HIGHLY DIVERGENT TANDEM REPEATS IN DNA SEQUENCES, STATISTICAL TESTS, AND BIOLOGICAL APPLICATION IN *DROSOPHILA MELANOGASTER* GENOME

***Boeva V.A.*[1], *Regnier M.*[2], *Makeev V.J.*[3]**

[1] Moscow State University, Vorob'evy Gory, Moscow, Russia; [2] INRIA Rocquencourt, Le Chesnay, France; [3]State Center GosNIIGenetika, Moscow, Russia
***** Corresponding author: e-mail: valeyo@csi.ru

Keywords: *divergent tandem repeats, statistical significance*

Summary

Motivation: Tandem repeats occur in regions of various functions in genomes. Highly divergent tandem repeats are difficult to separate from random background without focused statistical tests. Thus, the majority of tools designed to identify tandem repeats, which do not includes statistical evaluations of the results are not suitable for identification of divergent tandem repeats.

Results: We developed an effective algorithm for searching for highly divergent tandem repeats with evaluation of their statistical significance and implemented it in the SWAN software. In addition to identification of tandem repeats in a sequence, SWAN can be used to compute statistical significance of repeated structures identified by other popular tools designed for repeat finding. A number of biological examples of tandem repeats identified with SWAN are presented.

Availability: tool is available upon request from the authors.

Introduction

Repeated sequences often occur in prokaryotic and eukaryotic genomes. They are an example of the simplest regular structures in genome. A tandem repeat is a set of several similar words repeated contiguously in a sequence.

Tandem repeats differ in their 'fuzziness'. They can be '*strong*' or '*exact*' like ATTGC ATTGC and *approximate*, *divergent* or '*fuzzy*' like ATCGC ATGGC ATTCC. The size of the repeated unit is called a *period* and the number of copies is called an *exponent*.

Contiguously repeated patterns are often found in biological sequences, and sometimes are associated with functionally important segments. They could act as protein binding sites in enhancers and silencers, interact with transcription factors in promoters, affect the chromatin structure. In DNA coding regions long tandem repeats determine periodic structures in proteins. There are examples of repeats occurring in prokaryotic promoters.

Several tools have been published that identify tandem repeats in a nucleotide sequence. Among the best known are: **TROLL** (Castelo *et al*., 2002), **Tandem Repeat Finder** (Benson, 1999) and **mreps** (Kolpakov *et al*., 2003). However, these tools are mostly oriented to identify highly conserved repeated units.

The objective of our algorithm implemented as SWAN software is to identify series of repeated units without insertion or deletions, but with a high substitution rate. We consider only the repeated structures with a number of copies greater than two. Our algorithm identifies the length of the repeated unit and the number of repetitions. We have paid special effort to evaluate the statistical significance of the patterns found.

Algorithm

In this section we describe the main stages of the SWAN algorithm.

Detection of candidate tandem repeats is the starting stage of the algorithm. The objective of this stage is to identify all repeated patterns with period P and the number of mismatches between any three neighboring units less than some threshold K. Obviously, K varies between 0 and $2P$. The algorithm checks one by one all possible periods P from the user defined minimal value P_{min} to P_{max}. Here the parameter T, the average number of matches between neighboring units per letter, is important; $T = (2P - K)/P$ and takes the values between 0 and 2. Only words that have more than PT letters identical with the corresponding letters in the neighboring words are included into the pattern, with the letter i counted twice if it is identical with both letters $i+P$ and $i-P$. The candidate tandem repeats found may overlap, with the same segment of the sequence occupied by overlapping tandem repeats of different periods and exponents.

In the second stage, tandem repeats found in the first stage are processed. First the algorithm composes the IUPAC *consensus* for each of the repeated patterns. A set of words satisfying the consensus we call '*the motif*'.

The algorithm compares overlapping patterns and removes patterns whose units may be in turn presented as repeats with smaller period. In other words, if one repeat has a smaller period and another does not, only the second is retained. As a result, the remaining repeated units are not periodic.

To evaluate the statistical significance we formulated two probabilistic models. The first one is based on computation of the probability to find a motif repeated contiguously no less than n times (where n is the exponent) in a random sequence of length N given the condition that the motif has occurred at least once. This conditional probability literally reflects our searching algorithm: "for each word in the sequence one checks whether it is repeated n times". Is was developed using technique elaborated by M. Regnier (Regnier, 2002) and partially published in (Vandenbogaert, 2004). Minus logarithm of this probability is taken as *'Motif' Statistical significance*.

In the second statistical model the significance is evaluated based on the probability to find a structure as repeated pattern. In detail, for each tandem repeat one composes a '*mask*'. Let P be the period of the repeat and n be the exponent. 'Mask' is an intersection of P events. The i-event is the occurrence of the pre-defined number of identical letters simultaneously at positions i, $(i+P)$, $(i+2P)$, …, $(i+P(n-1))$. For example: for repeat ATC_ACG_AGC the 'mask' is the event that the same letter occurred three times at positions 1,4,7, then at positions 2,5,8 could be any three letters, and at least two identical letters occurred at positions 3, 6, 9. Minus logarithm of the probability of finding a 'mask' in the text is taken as *'Mask' Statistical significance*.

The last stage of the analysis of periodic patterns is the comparison of overlapping tandem repeats with elimination of less significant repeats. To this end, statistical significance criteria are used. It is possible to use statistical criteria derived from both 'motif probability' and 'mask probability.' If several repeats are detected in the same sequence region, only the repeat with the greatest statistical significance is kept and the others are filtered out.

Implementation

The SWAN software was written in C language, using some C++ features. The input of the program is a sequence file or a file containing sequence filenames and the following parameters: minimal similarity threshold between the repeated units; the minimal and maximal unit size limits; the model for statistical significance calculation (i.e. 'the motif model' or 'the mask model').

The program returns a single result file. It is a table containing the following information: the name of the sequence file; the start, the end, and the length of the repeat; the period size; the number of copies; the IUPAC consensus; the number of words in the motif that satisfy the consensus; 'the

motif probability' (*P*-value); 'the mask probability' (*P*-value); 'the motif statistical significance'; 'the mask statistical significance'; the tandem repeat itself. Table displays the output of the program.

Table. SWAN output

AR401352	2.24E-03	1008	1026	18	3	6	TKTTCC	2	5.09E-04	2.65
AR401358	1.01E-02	306	326	20	4	5	GDAVA	9	5.64E-03	2
AR430321	1.52E-02	921	942	21	3	7	TKRGRAA	8	7.26E-04	1.82

We included a special feature into our program that allows one to submit the output of other tandem repeat finders as the input to SWAN and evaluate the statistical significance supplying the necessary random model parameters.

The SWAN is available upon request from the authors.

Biological Application

The program was tested at several biological examples: *Drosophila melanogaster* locuses (16–120kb each) with marked exons (total length 49.4kb) and regulatory regions (total length 115.6kb), *Drosophila* heterochromatin sequences (under 40kb each) and arbitrary sequences from genome 'for comparison' (20kb-1Mb each).

Our program revealed that in various regions of locus (coding, regulatory, non-coding-or-regulatory regions) tandem repeats are distributed differently. It is shown in Fig.

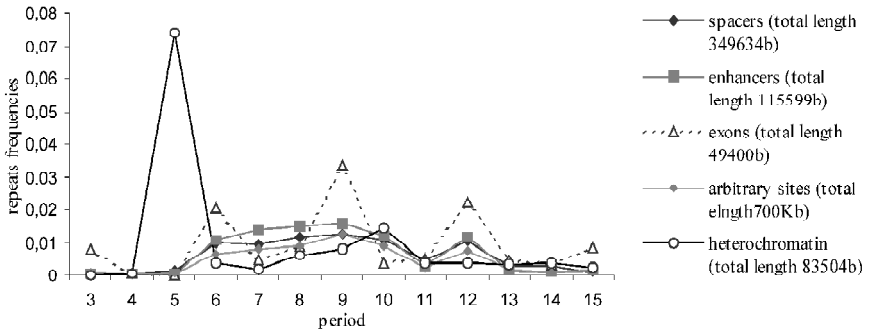


Fig. Frequencies for each period in various sites of genomes. Stat. significance e" 4. T=1.3.

It is noteworthy, that repeats with the length of repetitive unit divisible by three appear in coding regions much more frequently than in the other parts of genome. Obviously, it is related to the triplet structure of DNA code.

The majority of the tandem repeats in heterochromatin has periods multiple to 5, which is a half of the period of DNA helix pitch (about 10.2bp). In contrast, the majority of tandem repeats found in regulatory sequences has the period 7 or 8, which contradicts the DNA helix period. This may be related to the DNA packing in eu- and heterochromatin and DNA flexibility in euchromatin.

It is well-known that sometimes tandem repeats act as protein binding sites. In *Drosophila* an experimentally shown example of interaction with transcription factor *Bcd* is tandem repeat in *Otd*: CTCGGAT CTAGGAT CTTGCAT CTTGCAT CTCGCAT CTCGCAT. This repeat was found *de novo* by SWAN algorithm.

However, for the majority of repeated units found in *Drosophila* regulatory modules we could not find transcription factors involved into regulation of this system who bind to sites similar to the sequences of the repeated units. This question requires further studies.

Discussion

In this paper the algorithm was described that identifies tandem repeats in DNA sequences and evaluates their statistical significance.

We hope that the option of evaluation of statistical significance of the repeats found by other repeat finders would help to compare the results obtained by the other tools.

It is important that SWAN can identify repeats only with an integer exponent (the number of copies). At the same time TRF and **mreps** are able to find repeats with fractional exponents. But we do not think that this feature would restrict the application of our software, because our main objective is identification of highly divergent repeats with a limited period and a sufficiently high number of copies.

Acknowledgements

Authors are pleased to thank D. Papatsenko, G. Kravatskaya, N. Oparina, V. Ramensky and M. Vandenbogaert for discussion and data. This work has been supported by EcoNet 08159PG project and by RFBR grant 04-07-90270-в, and by Program in Molecular and Cellular Biology of Russian Academy of Sciences, coordinator V.G. Tumanyan. V. Makeev is also supported by CRDF RBO-1268-MO-02 and Howard Hughes Institute Grant 55000309.

References

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res., 1999. V. 27. P. 573–578.

Castelo A.T., Martins W., Gao G.R. TROLL–tandem repeat occurrence locator // Bioinformatics. 2002. V. 18. P. 634–636.

Kolpakov R., Bana G., Kucherov G. **mreps**: efficient and flexible detection of tandem repeats in DNA // Nucleic Acids Res. 2003. V. 31. P. 3672–3678.

Regnier M. Mathematical Tools for Regulatory Signals Extraction // Proc. of BGRS 2002. Novosibirsk: Russia, 2002. V. 1. P. 17–19.

Vandenbogaert M. Algorithmes and Mesures Statistiques pour la Recherche de Signaux Founctionels dans les Zones de Regulation. These de troisieme cycle de l'Universite de Bordeaux I. In French. L'Universite de Bordeaux I. 2004.

NON-RANDOM DISTRIBUTION OF ALU ELEMENTS IN HUMAN: NOVEL INSIGHTS FROM ANALYSIS OF THE COMPLETE GENOME

***Brahmachari S.K.*[1], *Grover D.*[1], *Majumder P.P.*[2], *Mukerji M.*[1]**

[1] Institute of Genomics and Integrative Biology, CSIR, Mall Road, Delhi, India; [2] Anthropology and Human Genetics Unit, Indian Statistical Institute, B.T. Road, Kolkata 700 108, India,
e-mail: skb@igib.res.in

Keywords: *Alu, repeat distribution, chromosome 21 and 22, functional classification, gene regulation, human genome, genomic composition*

Summary

Motivation: Comparison of human genome with sequences of other available genomes has revealed that the higher physiological and obvious observable complexity in humans cannot be explained just on the basis of coding sequences. Alu repeats, which occupy one-tenth of the human genome, have in the recent times shown to be involved in various aspects of gene regulation. In this study we have analyzed the distribution of Alu repeats in the complete genome in order to understand their functional significance in human.

Results: The highlights of the analysis are as follows (1) three-fourth of the total genes in the genome are associated with Alus (2) Alu density is higher in genes as compared to intergenic regions in all the chromosomes except 19 and 22 and large sized chromosomes were not necessarily associated with more Alu repeats. (3) Detailed analysis of Alus in chromosome 21 and 22 reveal a striking bias in distribution with respect to functional category of genes. (4) Genes involved in transport, metabolism and signaling are rich in Alu repeats whereas those of information pathway and structural proteins are not.

Introduction

Alu repeat elements belong to SINE (short interspersed nucleotide elements) family of repeats and are present predominantly in the non-coding regions of primate genomes. A majority of Alu repeats in human genome belong to old or intermediate subfamilies with relatively minimal representation of younger subfamilies (Deininger *et al*., 1992). Transcriptionally active regions of the genome are enriched in Alu elements (Korenberg, Rykowski, 1988). Earlier one of us has proposed that Alu like repeats can modulate gene expression, nucleosome organization and buffer conformation dependant regulation of gene expression (Conrad *et al*., 1986; Brahmachari *et al*., 1995). With the complete information about nucleotide sequence and genes now available publicly (International Human Genome sequencing Consortium, 2001), we have attempted in this study to explore various factors that may drive the integration of Alu repeats in the human genome.

Methods and Algorithms

The nucleotide sequence of human genome was downloaded from NCBI web site (ftp://ftp.ncbi.nlm.nih.gov/genomes/h_sapiens). Information about genes, introns and exons was extracted from the available data using custom-made PERL programs. Positions and subfamilies of Alu repeat in complete genome was identified using the program REPEATMASKER (http://repeatmasker.genome.washington.edu/) locally installed on compaq alpha sever ES40. Repeat numbers and density for each gene, intron, exon as well as chromosome was subsequently calculated using various PERL programs. Alu density in chromosomes and genes was expressed as Alu percentage, that is, percentage of the gene/chromosomal region occupied by Alus. For correlation analysis, each chromosome was split into 100 Kb intervals and Alu size, gene size,

intron size and GC content was calculated individually for all these regions. Statistical analysis was performed using Statsview package (ver 4.0).

The genes on chromosome 21 and 22 were classified into five functional classes namely structural proteins, information storage & processing, signaling pathways, metabolism and transport & binding proteins on the basis of information about function of the gene provided at various sites – Locus Link (http://www.ncbi.nlm.nih.gov/LocusLink/), GeneCard (http://bioinfo.weizmann.ac.il/cards/), Gene Quiz web server (http://www.sander.ebi.ac.uk/ gqsrv/), Gene Ontology (http:// www.geneontology.org) and UniGene database (http://www.ncbi.nlm. nih.gov/UniGene/). Only those genes, which were well characterized in terms of function and expression, were considered.

Statistical tests of significance and relationship among different variables, e.g. Alu subfamily frequencies, Alu percentage, functional class, chromosome and GC content, were carried out by chi square test, regression analysis and ANOVA.

Results

Alu distribution in whole genome. Alu repeats are present in 1,179,211 copies in the genome which together account for nearly 10.8 % of the sequenced region of human genome. Interestingly, large sized chromosomes were not necessarily associated with more Alu repeats (Fig. 1).

Alu repeat density and association with genes

Alu densities were found to be highly variable among different chromosomes in human, with chromosome 19 being the most Alu rich chromosome having 26.3 % of the region covered by these repeat elements followed by chromosomes 17 (19.1 %), 22 (18.2 %) and 16 (16.8 %). Chromosome Y, in addition to having the least number of Alu repeats, also has the least density (7.5 %). Statistical analysis revealed that Alu and gene densities are significantly correlated ($r = 0.83$, $p < 0.0001$). Chromosome 22 has the highest gene density in human genome (52.3 %) and chromosome Y has least (27.1 %).

Alus in intergenic and intragenic regions

Out of 27,963 genes so far identified in complete genome, Alu insertions were seen in nearly three-fourth of the genes. However, the density of Alu was non-uniform within genes (Fig. 2).

A comparative analysis between intragenic and intergenic regions showed that Alu coverage is higher in genes (12.5 %) than the intergenic regions (9.6 %) for all chromosomes except 19 and 22. Within genes, Alus were mostly seen in intronic regions, occupying 12.8 % of the intronic regions and were rarely found in exons (1.6 %).

Alu distribution in genes of different functional categories

We analyzed the Alu density in the genes of chromosomes 21 and 22 which with respect to five functional categories namely structural proteins, information storage & processing, metabolism, signaling pathway and transport & binding proteins. ANOVA of Alu density between the functional categories revealed that genes coding for structural proteins and information storage & processing components were either devoid or rarely associated with Alus. However, genes involved in metabolism and transport and binding processes and signaling were extremely rich in Alus (F-value = 14.294, $df = 4, 266$, $p < 0.0001$) (Fig. 3).

We also regressed out the effect of total GC content and GC contents and Alu percentages in the 25kb upstream and downstream regions had any significant effect on Alu percentage in genes of various functional categories and the results still remained significant (F ratio 14.314, $df = 4, 266$, $p < 0.0001$).

Discussion

Repetitive sequences, earlier considered non-functional, are being increasingly implicated in various regulatory functions and are believed to interact with the whole genome to influence its evolution. Our analysis on Alu repeats reveals many interesting features on their distribution in the human genome (Grover *et al*., 2003, 2004). Firstly, there is a significant correlation between Alu and gene densities across different human chromosomes indicating that gene density is a major driving factor for Alu accumulation within a chromosome. Moreover a higher Alu density in intra genic regions compared to intergenic regions indicates that these elements are preferred in genes. Though Alu density in the human genome was observed to be significantly influenced by gene density, intron density as well as GC content we found a very biased distribution of Alus in genes of different functional categories. Alus were clustered in genes involved in metabolic pathways, signalling and transport processes, whereas they were very poorly represented in genes coding for structural proteins and informational storage and processing components. Given the increasing evidence of involvement of Alus in various regulatory functions, it is intuitively obvious that they might be negatively selected in structural genes as well as the conserved information pathway genes. This non-random distribution of Alus is in agreement with the analysis of the first draft of the human genome wherein homeobox gene clusters, which are extremely conserved across evolution, are found to be devoid of or have low frequencies Alu elements.

Alu elements harbour binding sites for various tissue specific factors and hormone responsive elements, are involved in alternative splicing, can act as silencers as well as enhancers when present in 5'UTR as well as 3'UTR regions and also affect nucleosome positioning. The higher physiological complexity in primates compared to lower organisms has been attributed to considerable amount of changes in the metabolic machinery as well as transport mechanisms. Therefore, it is possible that these elements may be positively selected in genes involved in metabolism, transport and signaling processes because of a need for diverse regulatory functions in them. In contrast processes that are invariant among individuals like cell division, embryonic development, transcriptional and translational machinery are likely to be tightly regulated and are found to be Alu poor. It is also possible that higher Alu density in regulated genes may result in a higher number of epigenotypes, as subtle epigenetic variations can be brought about by these elements in a number of ways. Therefore, sequence analysis of repeat regions in human genome as well as phylogenetically related species is imperative for understanding the evolution of regulatory networks in homologous genes. With such wide spectrum of regulatory sites present in Alus as well as their involvement in nucleosome positioning, it also becomes imperative to screen for functional polymorphism in these sites, as they might be involved in the etiology as well as predisposition to many complex diseases which are triggered in response to hormonal imbalances as well as other environmental cues (Deininger, Batzer, 1999). A complex interplay of these elements could therefore be a dominant contributor to the human phenome.

Acknowledgments

Financial support from Dept. of Biotechnology (D.B.T.), Govt. of India on Programme on Functional Genomics to S.K.B. and Council of Scientific and Industrial Research (C.S.I.R.) is duly acknowledged.

References

Brahmachari S.K., Meera G. *et al*. Simple repetitive sequences in the genome: structure and functional significance // Electrophoresis. 1995. V. 16. P. 1705–1714.

Conrad M., Brahmachari S.K., Sasisekharan V. DNA structural variability as a factor in gene expression and evolution // Biosystems. 1986. V. 19. P. 123–126.

Deininger P.L., Batzer M.A. *et al*. Master genes in mammalian repetitive DNA amplification // Trends Genet.

1992. V. 8. P. 307–311.

Grover D., Majumder P.P. *et al*. Nonrandom distribution of alu elements in genes of various functional categories: insight from analysis of human chromosomes 21 and 22 // Mol. Biol. Evol. 2003. V. 20. P. 1420–1424.

Grover D., Mukerji M. *et al*. Alu repeat analysis in the complete human genome: trends and variations with respect to genomic composition // Bioinformatics. 2004, (in press).

Deininger P.L., Batzer M.A. Alu repeats and human disease // Mol. Genet. Metab. 1999. V. 67. P. 183–193

Korenberg J.R., Rykowski M.C. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands // Cell. 1988. V. 53. P. 391–400.

Lander E.S., Linton L.M. *et al*. Initial sequencing and analysis of the human genome // Nature. 2001. V. 409. P. 860–921.

MAPPING OF POTENTIALLY TRANSCRIBED REGIONS IN THE GENOME OF *E. COLI* BY NEW PROMOTER-SEARCH ALGORITHM

***Brok-Volchanski A.S.*[1], *Purtov Yu.A.*[1], *Lukyanov V.I.*[1], *Kostyanicina E.G.*[1], *Antipov S.S.*[1], *Deev A.A.*[2], *Ozoline O.N.**[1]**

[1] Institute of Cell Biophysics RAS, Pushchino, Moscow region, Russia; [2] Institute of Theoretical and Experimental Biophysics RAS, Pushchino, Moscow region, Russia
* Corresponding author: e-mail: ozoline@icb.psn.ru

Keywords: *Escherichia coli, promoter-search algorithm, transcription, regulatory regions, untranslated RNA*

Summary

Motivation: The total genome sequence of an organism should in principle encode the information that determines the regulatory networks controlling cellular metabolism. The understanding of the functioning of these networks both as independent and as connected integrated systems requires the identification of all the active components. In terms of gene expression an initial step is the mapping of promoter sites. Ideally this identification should be based on biochemical and genetic approaches. However current methods of promoter characterization are mostly attuned to the study of the regulation of particular genes rather than to identification of a whole set of functional promoters, which up to now is a complex task. Bioinformatics offers an alternative and powerful tool, which in combination with high-density oligonucleotide microarrays will be able to resolve this problem.

Results: Pattern-recognition software based on consensus elements and specific features of their genetic environments has been designed for identifying σ^{70}-specific promoters of *E. coli*. High predictive capability of this software estimated on the basis of testing compilations allowed reliable mapping of potentially transcribed regions within entire genome of this bacterium. Besides potential promoters in front of structural genes several hundreds of promoter-like regions have been found between convergently transcribed genes and within coding sequences. Most of them are expected as transcription start points for new genes. Here we discuss distribution of these sites according to known genes.

Availability: Compilation of known σ^{70}-specific promoters, as well as the map of known and predicted transcription start points are available by request (ozoline@icb.psn.ru).

Introduction

Genome-wide mapping of transcription regulatory sites is one of the key problems of genomic DNA annotation. Computer identification of these elements in front of known or predicted genes facilitates experimental analysis, while availability of the whole set of potentially transcribed regions allows revealing basically new genes. A large set of promoter-search algorithms has been proposed so far (reviewed in Hertz, Stormo, 1996; Horton, Kanehisa, 1992; see also Huerta, Collado-Vides, 2003). Most of them take into account two canonical hexamers near positions –35 and –12, some preferences in the regions flanking these elements and around start point of transcription. These algorithms are capable of identifying up to 90 % of known promoter sites but at this level even the best protocols recognize approximately 2 % of non-promoter DNA as a putative promoter-like sequences. Within sequences of genome size the background noise is, therefore, more than order of magnitude greater than the required signal. Taken into account this

intrinsic limitation, putative promoters for uncharacterized genes are searched within a narrow region (250 bp) from the experimentally estimated or predicted translation initiating codons. Using structural features in the genetic environment of consensus elements (Ozoline *et al.*, 2003) has increased selectivity of computer program. It allows identifying more than 85 % of natural promoters at the level when no false positives have been found in the control set. So high predictive capability gave a chance to reveal the whole set of potentially transcribed regions in the genomic DNA.

Results and Discussion

Approximately 91 % of known σ^{70}-promoters were identified when both strands of chromosomal DNA were scanned by newly designed software. Nearly 6 % of them appeared to be located in positions neighboring to the experimentally estimated transcription start point. Since accuracy of biochemical approaches allow 1–2 bp variations, we consider them as correctly recognized signals. Fig. 1 presents several typical examples.

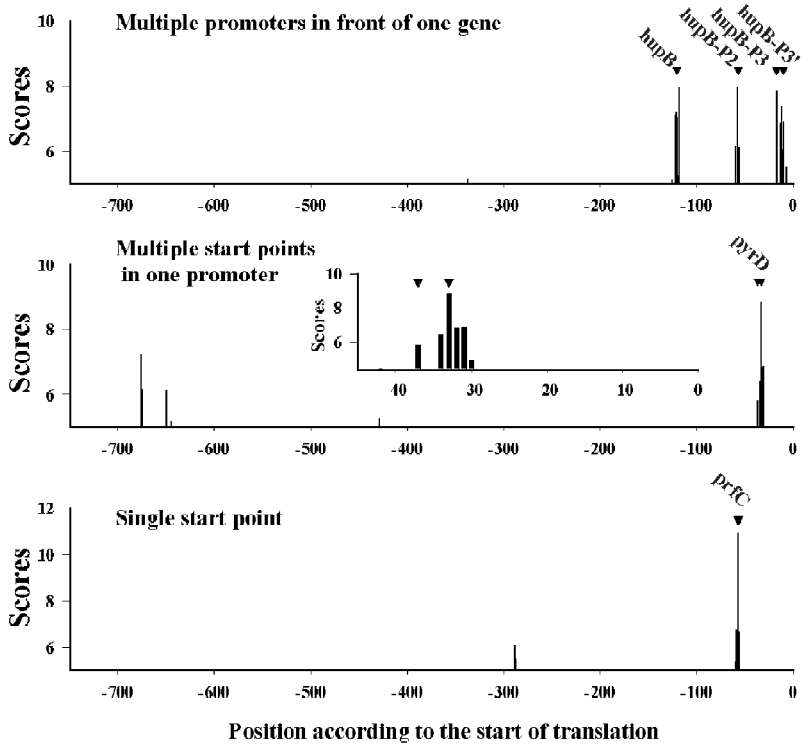


Fig. 1. Promoter-like scores within 750 bp regions upstream from translation initiation codons of indicated genes. Positions of known promoters are indicated by triangles. Inset on the central panel magnifies promoter region.

Nevertheless ~90 % of known promoters are located within 250 bp regions upstream from corresponding coding sequences (Fig. 1 and Fig. 2A), some genes have regulatory regions situated as far as 700–800 bp, thus estimating 750 bp as reasonable distance. Clusters of promoter-like sites (Fig. 1) (Huerta, Collado-Vides, 2003) often surround natural promoters. Positions with maximal scores in these clusters usually (>80 %) coincide with experimentally estimated start points, thus characterizing high predictive capability of our computer software.

Genome-wide distribution of promoter-like sequences. Only highly reliable signals ($p < 0{,}00005$) are used for this analysis. Only ~400 false positives are expected at this level. Putative promoters, composed in average of 20 promoter-like points have been found within 750 bp regions upstream from 2229 genes. Maximally scoring points have very similar with natural promoters positional distribution (Fig. 2A). Besides them, probable promoters have been found within other regions (Fig. 2B-E).

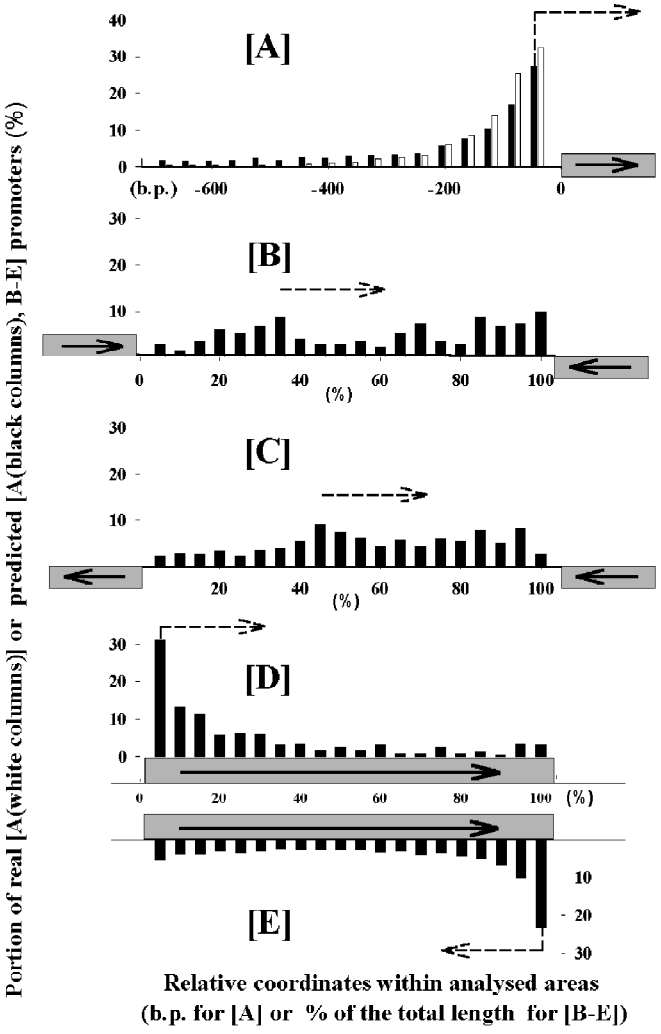


Fig. 2. Distribution of real and predicted promoters in respect to known genes (schematically indicated by rectangles). Solid-line arrows show directions of transcription. Dash-line arrows indicate orientation of predicted promoters. A - percentage of known (white) or predicted (black) promoters within 40 bp regions from initiating codon of translation. Putative promoters in other regions are sorted in 4 categories (B-E). To compare relative disposition of revealed promoters according to the schematically shown genes, all analyzed areas were subdivided into twenty 5 % regions.

Putative promoters in intergenic regions. More than 600 promoters have been found between genes transcribed convergently (Fig. 2B) or from another strand (Fig. 2C). An average distance between non-overlapping genes in the genome of *E. coli* is 148 bp. An average length of intergenic regions containing potential promoters is larger (440 and 298 for sets of Fig. 2B and 2C, respectively), thus providing a possibility for the existence of additional genes. They may code short polypeptides or untranslated RNAs. Thus, genes of 109 b. regulatory RNA spf and 377 b. RNA rnpB are located between convergently transcribed genes. A possibility for alternative transcription of downstream genes also should be taken into account. Distribution of revealed promoter-like signals does not show any preference in positioning (Fig. 2C, B). Their appearance, therefore, could not be explained by artifacts caused by the presence of regulatory regions associated with ends of neighboring genes.

Putative promoters within coding regions. Potential promoters with the same as working promoter orientation have been identified within translated regions of 379 genes (Fig. 2D). Even though located at longer than 750 bp distances, many of them may intensify expression of the next genes in operons. Well-expressed preference in the positioning of these promoter-like sequences in the beginning of translated regions allows, however, a possibility that some of them are only part of clusters encircling normally located promoters and appear due to the specific structural features in the transcription regulatory regions.

Potential promoters for anti-sense transcription. There are at least 709 genes containing strong promoter-like signals with propensity to transcribe anti-sense RNAs (Fig. 2E). Forming mRNA-RNA duplexes such products may affect efficiency of translation. *E. coli* contains at least 2 such RNAs (CopA and SsrA). Expected locations (beginning of the main gene) have, however, only ~20 % of revealed promoters, thus requiring another explanation for the majority of these signals. We found that ~17 % of genes from this set have intrinsic promoters with both orientations and within 67 genes there is a possibility to produce mutually complementary RNAs with an average length of 619 bases. As far as we know such products are not registered in *E. coli*. Genome-wide mapping of potentially transcribed sites revealed, therefore, some unexpected features in their relative distribution.

Acknowledgements

The studies are supported by the RFBR(03-04-48339) and RFBR-naukograd(04-04-97280).

References

Hertz G.Z., Stormo G.D. Escherichia coli promoter sequences: analysis and prediction // Meth. Enzymol. 1996. V. 273. P. 30–42.

Horton P.B., Kanehisa M. An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites // Nucl. Acids Res. 1992. V. 20. P. 4331–4338.

Huerta A.M., Collado-Vides J. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals // J. Mol. Biol. 2003. V. 333. P. 261–278.

Ozoline O.N., Brok-Volchansky A.S., Deev A.A. Promoter-search algorithm based on canonical and non-canonical sequence elements of *E. coli* regulatory regions // J. Biomol. Struct. and Dynam. 2003. V. 20. P. 905–906.

ACCURATE PREDICTION OF DNA OPENING PROFILES BY PEYRARD-BISHOP NONLINEAR DYNAMIC SIMULATIONS

***Choi C.H.*[1], *Kalosakas G.*[2], *Rasmussen K.O.*[2], *Bishop A.R.*[2], *Usheva A.**[1]**

[1] Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA; [2]Los Alamos National Laboratory, Los Alamos, NM, USA
* Corresponding author: e-mail: ausheva@bidmc.harvard.edu

Keywords: *physical properties of DNA, thermal opening profiles, DNA dynamics, S1 nuclease*

Summary

Motivation: The Peyrard-Bishop nonlinear model has proven to be an accurate predictor of the elastic properties of DNA. Through dynamical simulations, it is possible to gather statistical data on the local opening propensity of a limited sequence of double-stranded DNA. It is important to compare these computational results with experiment, to evaluate the usefulness of the Peyrard-Bishop model in a predictive capacity.

Results: Simulation and analysis of three linear DNA duplexes yields three distinct opening profiles by the Peyrard-Bishop model. Controlled digestion of radioactively-labeled templates by S1 nuclease, which selectively cleaves single-stranded DNA, shows an excellent correlation of the predicted openings with experimental data.

Introduction

It has long been known that double-stranded DNA is subject to temporary, localized openings of its two strands (Levitt, 1983). The available thermal energy in the system serves to destabilize the double helix structure at specific positions on the DNA sequence. These localized openings are significant enough in both size and duration to allow for chemical reactions to occur. NMR proton exchange measurements show that these openings may have lifetimes as long as 1 ms (Gueron, Leroy, 1995; Leroy *et al*., 1988). Theoretical studies have also supported this data, predicting significant distortions in the equilibrium structure (Giudice, Lavery, 2003; Levitt, 1983). These openings have been described in the physical literature as solitons or в-premeltons, and it has been speculated that they may play a role in the sequence-dependent opening of double-stranded DNA and transcription (Banerjee, Sobell, 1983; Prohofsky, 1988). As these dynamic openings may be functionally important sites in DNA, it is advantageous to develop experimental assays and computer models which can recreate "opening profiles" for DNA fragments. Here we report the close correspondence of S1 nuclease cleavage assay results with highly accurate Peyrard-Bishop model dynamic simulations (Peyrard, Bishop, 1989).

Model and Methods

S1 nuclease cleavage assays. S1 nuclease is a member of a family of glycoprotein nucleases which has been shown to selectively cleave single-stranded DNA (Wiegand *et al*., 1975). It has been widely employed in studies of transcriptional systems. The bulkiness of S1 nuclease aids in selectively cleaving larger temporary openings over small openings, but it results in a very weak signal. We overcame this issue of sensitivity by incubating the dsDNA with S1 nuclease for a longer period of time than in usual assays (45 mins). The noncoding (lower) strand for the non-promoter control and the P5 promoter, and the coding (upper) strand for the AdMLP was [^{32}P]-labeled at the 5'-terminus with T4 polynucleotide kinase (Invitrogen). The labeled strand and the unlabeled strand were then annealed by temperature cycling, and the dsDNA was used as a substrate for S1 nuclease cleavage. 0.2 nM dsDNA was incubated with S1 nuclease (50 units of enzyme

per reaction) at 28 °C for 45 minutes in buffer containing 10 mM HEPES (pH 7.2), 50 mM NaCl and 4 mM $Zn(C_2H_3O_2)_2$, as recommended by the supplier (Roche). The reaction was stopped with 20 mM EDTA. After ethanol precipitation, the DNA digestion products were electrophoretically separated on a 10 % sequencing gel (National Diagnostics). A Molecular Dynamics Phosphorimager 400-B was used to document the results.

Model. A modified Peyrard-Bishop model (Dauxois *et al*., 1993) with optimized parameters was applied in Langevin dynamical simulations. The sequence was repeated on both ends of the fragment to avoid terminal base pair effects, effectively circularizing the DNA sequence without any torsional effects. Simulations were run on several Sun Blade workstations. 100 separate PB model realizations were simulated over a 1 ns timescale using 1 femtosecond intervals, yielding 10^8 data points per sequence. Instances of a 10 bp opening over a threshold value of 2.1 Angstroms were recorded, and the values at the central base pair are reported.

Implementation and Results

We have recently demonstrated that destabilized regions of double-stranded DNA fragments are sensitive to S1 nuclease digestion under mild conditions, and that opening profiles can be derived from these assays (Choi *et al*., 2004). We chose four DNA sequences to examine for our studies (Table). The S1 nuclease assay results vary with the DNA sequences tested (Fig. 1).

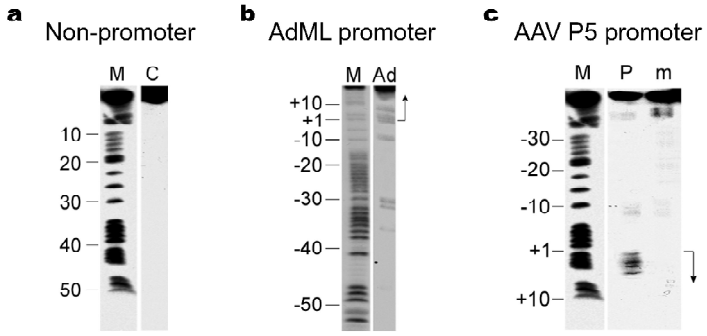


Fig. 1. S1 nuclease cleavage assays of four tested DNA sequences. (a) 62 bp sequence randomly selected from the cDNA of human transcription factor IIB; (b) 86 bp sequence of the adenovirus major late (AdML) promoter; (c) 69 bp sequence of the adeno-associated viral (AAV) P5 promoter and an associated mutant sequence. Numbered base pair positions are labeled to the left of each panel, and the associated DNA marker ladders are shown in lane M.

Table. Experimental sequences tested

Name	Sequence
62 bp sequence from cDNA of hTFIIB	CATATAGCCCGTAAAGCTGTGGAATTGGACTTGGTTCCTGGGAG GAGCCCCATCTCTGTGGC
86 bp AdML promoter	GCCACGTGACCAGGGGTCCCCGCCGGGGGGGTATAAAAGGGGG CGGACCTCTGTTCGTCCTCACTGTCTTCCGGATCGCTGTCCAG
69 bp AAV P5 promoter	GTGGCCATTTAGGGTATATATGGCCGAGTGAGCGAGCAGGATCT CCATTTTGACCGCGAAATTTGAACG
69 bp AAV P5 mutant promoter	GTGGCCATTTAGGGTATATATGGCCGAGTGAGCGAGCAGGATCT CCGCTTTGACCGCGAAATTTGAACG

Molecular dynamics simulations were run with these same DNA sequences, and the results compiled into a sequence-dependent opening profile. Comparison of the Peyrard-Bishop (PB) model-predicted opening profiles of the DNA fragments with the density profiles from these experimental assays shows a clear correspondence (Fig. 2).

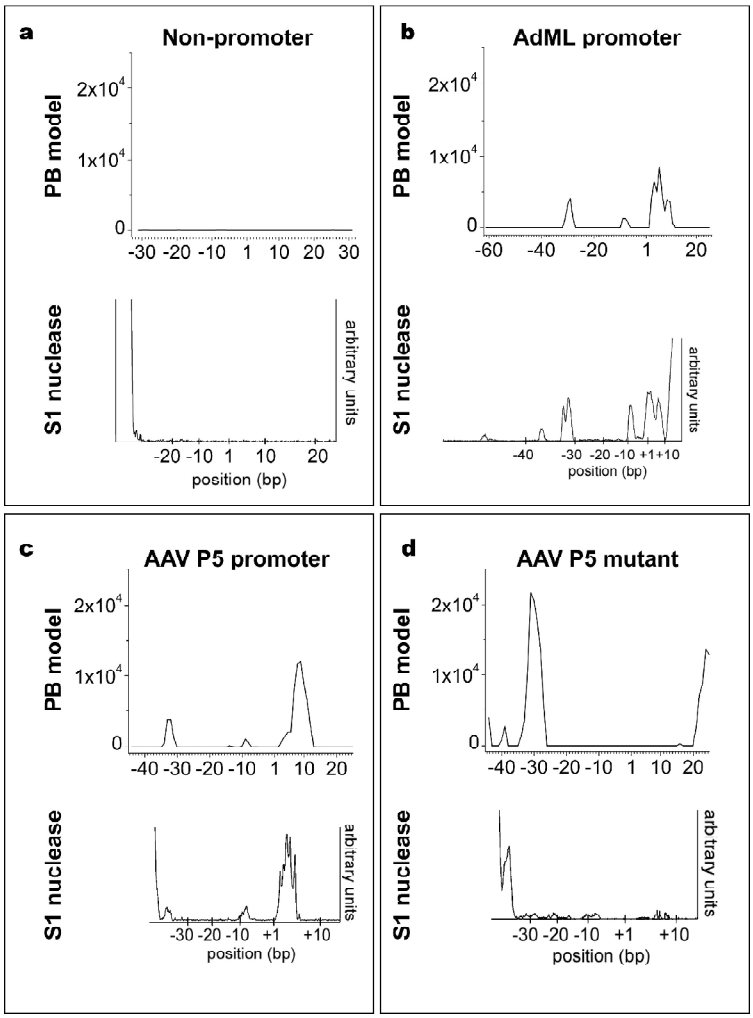


Fig. 2. Close correspondence between PB model predictions and S1 nuclease experimental assays. PB model opening profiles plot base pair position versus recorded instances of large opening formation in simulations. (a) 62 bp non-promoter sequence; (b) 86 bp AdML promoter; (c) 69 bp AAV P5 promoter. (d) 69 bp AAV P5 mutant promoter.

Discussion

The Peyrard-Bishop model has been shown to accurately describe the denaturation of DNA, and experimental results support its predictions (Campa, Giansanti, 1998). We report that even for oligonucleotides of around 100 bp in length, the method precisely simulates DNA dynamical motion, recreating thermal fluctuational openings which are also observed through experiments with S1 nuclease. We conclude that the Peyrard-Bishop model stands as a valuable tool for the prediction of these phenomena in double-stranded DNA.

Acknowledgements

This work was supported by NIH grant HL624458 to A.U. and by the US Department of Energy under contract W-7405-ENG-36.

References

Banerjee A., Sobell H. M. Presence of nonlinear excitations in DNA structure and their relationship to DNA premelting and to drug intercalation // J. Biomol. Struct. Dyn. 1983. V. 1. P. 253–262.

Campa A., Giansanti A. Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains // Physical Review E Statistical Physics, Plasmas Fluids, Related Interdisciplinary Topics. 1998. V. 58. P. 3585–3588.

Choi C.H., Kalosakas G., Rasmussen K., Hiromura M., Bishop A.R., Usheva A. DNA dynamically directs its own transcription initiation // Nucleic Acids Res. 2004. V. 32. P. 1584–1590.

Dauxois T., Peyrard M., Bishop A.R. Dynamics and thermodynamics of a nonlinear model for DNA denaturation // Physical Review E Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics. 1993. V. 47. P. 684–695.

Giudice E., Lavery R. Nucleic acid base pair dynamics: the impact of sequence and structure using free-energy calculations // J. Am. Chem. Soc. 2003. V. 125. P. 4998–4999.

Gueron M., Leroy J.L. Studies of base pair kinetics by NMR measurement of proton exchange // Methods Enzymol. 1995. V. 261. P. 383–413.

Leroy J.L., Charretier E., Kochoyan M., Gueron M. Evidence from base-pair kinetics for two types of adenine tract structures in solution: their relation to DNA curvature // Biochemistry. 1988. V. 27. P. 8894–8898.

Levitt M. Computer simulation of DNA double-helix dynamics // Cold Spring Harb Symp Quant Biol. 1983. V. 47. Pt 1. P. 251–262.

Peyrard M., Bishop A.R. Statistical mechanics of a nonlinear model for DNA denaturation // Physical Review Letters. 1989. V. 62. P. 2755–2758.

Prohofsky E.W. Solitons hiding in DNA and their possible significance in RNA transcription // Physical Review A. 1988. V. 38. P. 1538–1541.

Wiegand R.C., Godson G.N., Radding C.M. Specificity of the S1 nuclease from Aspergillus oryzae // J. Biol. Chem. 1975. V. 250. P. 8848–8855.

DEVELOPMENT OF A METHOD FOR *IN SILICO* IDENTIFICATION OF DNA SEQUENCES PARTICIPATING IN MEIOTIC CHROMOSOME SYNAPSIS AND RECOMBINATION

***Dadashev S.Ya., Grishaeva T.M., Bogdanov Yu.F.* ***

N.I. Vavilov Institute of General Genetics RAS, Moscow, Russia
* Corresponding author: e-mail: bogdanov@vigg.ru

Keywords: *chromosome, DNA, meiosis, recombination, primary structure, higher levels, repetitive sequences, synapsis, synaptonemal complex, Y chromosome*

Abstract

Motivation: Eukaryotic chromosomes are organized in loops. Bases of chromatin loops are attached to chromosome axes; in meiosis, these are protein axes forming the synaptonemal complex (SC). Earlier, SC attachment regions (SCAR DNA) were isolated and characterized in rat and hamster. Their sequences were shown to form a specific family (Karpova *et al*., 1995). Hence the problem was to develop a method of *in silico* searching for specific SCAR DNA-like meiotic sequences (mei-DNA) in genomes of other organisms. Detection of mei-DNA would pose the following questions: (1) whether mei-DNA plays a part in the formation and/or topography of lateral chromatin loops and (2) how is mei-DNA related to sites of meiotic recombination.

Results: A method was developed to *in silico* search the mammalian genomes for specific repetitive sequences (RS) similar to SCAR DNA. The chromosome distribution of these (mei-DNA) sequences was characterized. The frequency of mei-DNA sites in the Y chromosome is almost thrice lower than in autosomes in human. Autosomal mei-DNA sites are mostly 10 or 50–60 kb apart, suggesting clustering. The meiotic recombination frequency is minimal in sites of predominant mei-DNA location. The results agree with the fact that most of the Y chromosome is not involved in the SC formation or meiotic recombination and with the model implying that mei-DNA sequences contact the SC central element and lie at the bases of lateral chromatin loops, while recombination initiation sites are more often located in central regions of the loops.

Introduction

Chromosome axes are formed by linear assembly of chromatin loops, which occur at an evolutionarily conserved density of ~ 20 loops per micron of axis length (Zickler, Kleckner, 1999). In meiotic prophase, the axes form morphologically detectable axial cores (AC) as a result of incorporation of meiosis-specific cohesin Rec8 and major proteins SCP2 and SCP3 (in mammals). Pairwise association of AC of homologous chromosomes leads to the formation of synaptonemal complexes (SC).

A specific DNA fraction corresponding to SC attachment regions (SCAR DNA) was isolated experimentally. SCAR DNA sequences share a set of features and may consequently be detached in a specific family (Karpova *et al*., 1989; Pearlman *et al*., 1992).

As SC formation starts, meiotic recombination is simultaneously initiated via programmed double-strand breaks (DSBs) (Keeney, 2001; Lichten, 2001). Many recombinational interactions are initiated along each chromosome, but only a few finally mature into crossovers. This takes place after a discrete series of biochemical steps (Kleckner *et al*., 2003).

We developed a method to *in silico* search the mammalian genomes for DNA sequences involved in homologous chromosome synapsis and meiotic recombination (mei-DNA).
The questions are natural as to (1) whether mei-DNA takes a part in the formation and/or topography of lateral chromatin loops and (2) what are the relationships between mei-DNA and sites of DSBs or realized meiotic recombination (crossing over)?

Methods

We used the Human Genome Resources database (http://www.ncbi.nlm.nih.gov/genome/guide/human/) and the program Blast the Human Genome (http://www.ncbi.nlm.nih.gov/BLAST/). Mei-DNA sequences were identified as follows. A sequence under study was fed into the Blast program by 6000 nt with a 5000-nt periodicity. Search parameters: Megablast off, Expect 10, Filter by default, Descriptions 10000, and Alignments 0. When the total number of similar sequences found exceeded 268, the portion of Y-chromosomal sequences was established. Statistical analysis was performed with the STATISTICA v. 6 program (StatSoft, Inc., 2001) (STATISTICA data analysis software system, version 6, www.statsoft.com.).

Implications and Results

To identify the DNA sequences involved in chromosome synapsis and meiotic recombination, advantage was taken of the fact that the Y chromosome differs in meiotic behavior from autosomes in mammals. SC between the X and Y chromosomes is formed only in a limited chromosome region, which accounts for about 10 % of the Y chromosome length in human. These observations indicate that the Y chromosome contains a few DNA sequences responsible for the attachment of chromatin filaments to SC. These sequences are hereafter referred to as meiotic DNA (mei-DNA). Only short (300- to 1000-bp) mei-DNA fragments were isolated experimentally and termed SCAR DNA (Karpova *et al*., 1989; Pearlman *et al*., 1992).

Chromatin loops are probably attached to SC via DNA-protein interactions. Since the interactions are likely uniform, at least to some extent, throughout the SC length, we assumed that the relevant DNA sequences are repetitive (RS).

Programs Blast the Mouse Genome and Blast the Human Genome (http://www.ncbi.nlm.nih.gov/BLAST/) reveal sequences homologous to a reference one and localize them on chromosomes. Basing on the putative function of mei-DNA, we used the Blast program to identify the mei-DNA sequences in the human genome. We expected mei-DNA to occur in multiple copies (300–1000 bp each) in all but one, Y, chromosome.

1. A method of searching for mei-DNA was developed with the Blast the Mouse Genome program and samples of RS extracted from GenBank. These RS were annotated as (1) golden hamster and rat DNAs tightly associated with SC (SCAR DNA) (Karpova *et al*., 1995; Pearlman *et al*., 1992), (2) mouse DNA sequences associated with the nuclear lamina (lamina DNA) (Christova *et al*., 1992), (3) mouse DNA sequences associated with the core of rosette-like structures in interphase chromatin (CRLS DNA) (Glazkov *et al*., 1994), and (4) chicken DNA sequences associated with the nuclear matrix (SAR/MAR DNA) (Krajewski *et al*., 1992). These DNA sequences are all similar in size and in copy number per genome and are classed with moderate repeats. We estimated the Y-chromosomal portion in the total number of RS of a given class (Table).

Table. Copy number of DNA families isolated from nuclear ultrastructures, found in the total genome and in the Y chromosome

Repetitive sequence family	SCAR DNA	Lamina DNA	CRLS DNA	SAR/MAR DNA
Total copy number	3349	4658	2220	1708
Y-chromosomal copy number	3	12	27	4
Portion of Y-chromosomal sequences	0.0009	0.0026	0.012	0.0023

In the case of SCAR DNA, the portion was significantly lower than in the cases of lamina DNA ($p < 0.042$) and CRLS DNA ($p < 0.0001$) and lower, though nonsignificantly, than in the case of SAR/MAR DNA. This agrees with the known fact that most of the Y chromosome is neither involved in SC nor subject to recombination. Thus, our assumption that the Y chromosome is poor in mei-DNA was confirmed by the example of well-known SCAR DNA. This feature may be used to identify the mei-DNA sequences. Hence we obtained a means of *in silico* searching for the sequences of interest in various genomes.

2. Further search was performed in the human major histocompatibility complex (MHC), for which a detailed recombination map has been constructed (Cullen *et al*., 2002). We observed that it is possible to isolate at least three RS classes differing in representation in the Y chromosome. The coefficients of correlation were calculated for the recombination frequency and the density of RS distribution in MHC for each class. Sequences that are rare in the Y chromosome were of interest. We found that some RS are close to recombination hot spots ($r = 0.3$ at $p < 0.11$), whereas some others avoid these ($r = -0.44$ at $p < 0.02$).

3. To identify the RS class harboring mei-DNA, we analyzed the chromosome distribution of RS with the example of an arbitrarily selected 1.5-Mb DNA sequence of human chromosome 2 (gi 29789875:612597-2112597 H s chr2 genomic contig NT_022135.13). For each RS class, we estimated the copy number, the portion of Y-chromosomal copies, and the interval between copies. A sequence was assigned to mei-DNA when the copy number of the relevant RS class exceeded 268 and the Y-chromosomal portion of these RS was less than 0.003. Copy numbers beyond 268 allow reliable detection of a lower representation of relevant RS in the Y chromosome (by analogy with SCAR-DNA). The other parameter (0.003) was selected on the basis of between-copy intervals established for each RS class. The RS class with a portion of Y-chromosomal sequences less than 0.003 (0.3 %) best fits the model of chromatin organization in loops around SC. The most common distances between RS copies of between-copy intervals in this class (mei-DNA) as it follows from frequency histogram (not shown) are 10 and 50-60 kb, suggesting clusters located 50-60 kb apart. This agrees with the model of meiotic chromatin organization in loops, assuming that mei-DNA is close to loop bases and participates in synapsis of homologs.

4. We studied the relationships between mei-DNA and recombination with the example of human MHC. The frequency distribution of recombination sites in MHC was collated with the distribution of mei-DNA sequences (not shown here). It is clearly seen that the recombination frequency is minimal in regions of predominant mei-DNA location and vice versa. This conclusion was supported by Spearman's correlation coefficient calculated for these parameters ($r = –0.46$ at $p < 0.02$; after smoothing, $r = –0.72$ at $p < 0.00002$).

5. According to the model of meiotic chromosome structure (Kleckner *et al*., 2003), only a single recombination event may be initiated in one chromatin loop. Then, the higher the number of loops in a given DNA region, the higher the frequency of recombination. Figure 3 shows the recombination frequency as dependent on the copy number of mei-DNA sequences having the portion of Y-chromosomal copies less than 0.003. At a low mei-DNA density, the recombination frequency shows a positive linear dependence on this parameter. After a maximum (two mei-DNA copies per 100 kb, which corresponds to a loop size of 50 kb), the dependence becomes negative, and then recombination frequency drops to a minimal level independent of the mei-DNA density.

To explain the decrease in recombination frequency with decreasing loop size, it is possible to assume that a contact is difficult between short homologous chromosome loops attached to opposite lateral SC elements, while such contacts are essential for recombination initiation. This fact was not taken into account by the authors of the model. Our data make it possible to check another assumption implied by the model, namely, the one loop-one recombination event rule. We normalized the above data to obtain the recombination frequency per mei-DNA copy (i.e., per loop according

to our views). If the above rule were true, the recombination frequency per loop would be independent of the loop size. However, the number of recombination events per loop decreases with increasing mei-DNA copy number and, respectively, with decreasing loop size. This indicates that recombination depends on the size of the "accessible" (Kleckner *et al.*, 2003) loop region rather than on the number of loops.

Discussion

Our results obtained with the method of *in silico* identification of structural mei-DNA sequences agree with experimental data and theoretical views of the organization of meiotic chromosomes. Mei-DNA is involved in the organization of lateral chromatin loops but not in recombination initiation. Supported by Russian Foundation for Basic Research, project #02-04-48761.

References

Christova R., Bach I., Galcheva-Gargova Z. Sequences of DNA fragments contacting the nuclear lamina *in vivo* // DNA Cell Biol. 1992. V. 11. P. 627–636.

Cullen M., Perfetto S.P, Klitz W., Nelson G., Carrington M. High-resolution patterns of meiotic recombination across the human major histocompatibility complex // Am. J. Hum Genet. 2002. V. 71. P. 759–776.

Glazkov M.V., Poltaraus A.B., Lebedeva I.A. Nucleotide sequence of DNA isolated from protein cores of rosette-like structures (elementary chromomeres) of mouse interphase chromosomes // Genetika (Moscow). 1994. V. 30. P. 1146–1154.

Karpova O.I., Penkina M.V., Dadashev S.Y., Mil'shina N.V., Hernandes J., Radchenko I.V., Bogdanov Yu.F. Features of the primary structure of DNA from the synaptonemal complex of the golden hamster // Mol. Biol. (Mosk). 1995. V. 29. P. 512–521.

Karpova O.I., Safronov V.V., Zaitseva S.P., Bogdanov Yu.F. Various properties of DNA from isolated fractions of the mouse synaptonemal complexes // Mol Biol (Mosk). 1989. V. 23. P. 571–579.

Keeney S. Mechanism and control of meiotic recombination initiation // Curr. Top, Dev. Biol. 2001. V. 52. P. 1–53.

Kleckner N., Storlazzi A., Zickler D. Coordinate variation in meiotic pachytene SC length and total crossover/chiasma frequency under conditions of constant DNA length // Trends Genet. 2003. V. 19. P. 623–628.

Krajewski W.A., Razin S.V. Organization of specific DNA sequence elements in the region of the replication origin and matrix attachment site in the chicken alpha-globin gene domain // Mol. Gen. Genet. 1992. V. 235. P. 381–388.

Lichten M. Meiotic recombination: breaking the genome to save it // Curr. Biol. 2001. V. 11. R253–R256.

Pearlman R.E., Tsao N., Moens P.B. Synaptonemal complexes from DNase-treated rat pachytene chromosomes contain (GT)n and LINE/SINE sequences // Genetics. 1992. V. 130. P. 865–872.

Zickler D., Klecklner N. Meiotic chromosomes: integrating structure and function // Annu. Rev. Genet. 1999. V. 33. P. 603–754.

CONSERVATION OF ALTERNATIVE SPLICING REGULATORY SIGNAL UGCAUG IN THE MOUSE AND HUMAN GENOMES

***Denisov S.V.*[1], *Gelfand M.S.**[2,3]**

[1] M.V. Lomonosov Moscow State University, Moscow, Russia; [2] Institute for Information Transmission Problems, RAS, Moscow, Russia; [3] State Scientific Center GosNIIGenetika, Moscow, Russia
* Corresponding author: e-mail: gelfand@iitp.ru

Keywords: *alternative splicing, regulation, human genome*

Resume

Motivation: It was previously shown that the motif UGCAUG is significantly over-represented in those downstream of brain-specific cassette exons as compared to introns downstream of constitutive exons (Brudno *et al*., 2001). Thus it is of interest to study the conservation of these UGCAUG-sites in the mouse and human genomes in order to see whether they are functional or not.

Results: We analyzed the conservation of all sites UGCAUG that occurred in intron sequences downstream of cassette exons (Brudno *et al*., 2001); 50 % of human sites and 70 % of mouse sites areconserved in both genomes. Because functional regions of the genome (e.g. protein-coding sequences, regulatory signals) evolve significantly slower than the non-functional, the conserved sites are likely functional and are regulatory sites of alternative splicing.

Introduction

The 75 million years that elapsed since the divergence of the human and mouse lineages led to a substantial divergence in neutral DNA, whereas the constraint on functional elements kept them conserved. Indeed, orthologous human and mouse exons are, on average, 85 % identical, but introns are less conserved: 60 % of the nonexonic sequences are nonalignable, and the average identity level in the alignable regions is 69 % (Waterson *et al*., 2002). However, regions that are conserved between the human and mouse genomes are also found in introns (e.g. Sorec, Ast, 2003). These regions may be important to the regulation of alternative splicing. Brudno *et al*. (2001) analyzed a group of brain-specific cassette exons and adjacent introns. It was demonstrated that the UGCAUG motif was over-represented in downstream regions of alternative brain-specific exons (the control group consisted of downstream regions of constitutive exons). Sorec and Ast (2003) found that 77 % of conserved alternatively spliced exons were flanked on both sides by long conserved intronic sequences. In comparison, only 17 % of conserved constitutively spliced exons were flanked by such conserved intronic sequences. The most abundant hexamer in the conserved intronic sequences downstream of the alternatively spliced exons was UGCAUG. At the same time, this hexamer was not over-abundant downstream of constitutively spliced exons, nor in intronic sequences located upstream of alternatively spliced exons. There is experimental evidence that the KSRP and FBP proteins bind directly to the UGCAUG hexanucleotide (Ladd, Cooper, 2002).

Methods and Algorithms

For the sample of exons from (Brudno *et al*., 2001) we analyzed the conservation of UGCAUG sites between the human and mouse genomes and studied the specificity of tissue and organs expression of cassette exons.

Analysis of conservation involved the following steps: (a) identification of orthologous genes; (b) identification of homologous alternative exons; (c) finding all UGCAUG motifs in proximal downstream intronic sequences (the length of the analyzed sequence was 1000 nt) for mouse and human separately; (d) alignment of the mouse and human intronic sequences, which contained UGCAUG.

Implementation and Results

We found 28 UGCAUG sites in the human intronic sequences and 20 sites in the mouse. The data on site conservation are given in Table; 14 motifs are conserved in both genomes. Thus, the conservation level is 14/25 H" 50 % for human and 14/20 = 70 % for mouse.

Previously, all exons from this sample were believed to be brain-specific. We analyzed tissue specificity of the exon expression using EST/mRNA data and found that at least 10 human exons (Nos. 2,3,5,7,8,11,13,17,19,25 in Table) and 4 mouse exons (Nos. 2,8,10,24 in Table) are not brain-specific. Thus, UGCAUG is not a brain-specific enhancer of alternative splicing.

Table. Conservation of UGCAUG sites between mouse and human

No	Gene name / synonym	UGCAUG position (intron starts at pos. 1)		Alignment	True or false signal?
		Human	Mouse		
01	Ankyrin B, large insert / ANK2	-------	1. 226	exists	false
02	FHL1B /FHL1				
03	PMCA4 calcium pump / ATP2B2	1. 22	-------	does not exist	false
04	SCN8 sodium channel / SCN8A	1. 30	1. 30	exists	true
05	Amphiphysin II (region I) / BIN1	-------	-------		
06	N-type Ca channel / CACNA1B	-------	-------		
07	NMDA-R1 exon 5 / GRIN1	1. 8	1. 8	exists	true
		2. 257	2. 262	exists	true
		3. 494	3. 497	exists	true
08	CLCB / CLTB	-------	-------		
09	Myelin-associated glycoprotein exon 12 / MAG	1. 63	1. 63	exists	true
			2. 599	does not exist	false
10	4.1R exon 15 / EPB41	-------	-------		
11	B-KSR1 / KSR	1. 12	-------	exists	?
12	4.1N / EPB41L1	1. 5	1. 5	exists	true
13	4.1B exon 15 / EPB41L3	1. 286	1. 271	exists	true
		2. 344	2. 328	exists	true
14	HDlg / DLG1	1. 243	1. 243	exists	true
		2. 275	2. 274	exists	true
15	KOR-3a / OPRK1		1. 304	excluded	
16	Agrin exon 33 / AGRN	-------	-------		
17	MHC-B / MYH10	1. 244	-------	does not exist	false
18	NF1 exon 9a / NF1	1. 57	1. 57	exists	true
		2. 117	2. 118	exists	true
19	LAR tyrosine phosphatase / PTPRF	1. 884	1. 912	does not exist	false
		2. 888	-------	does not exist	false

No	Gene name / synonym	UGCAUG position (intron starts at pos. 1)		Alignment	True or false signal?
		Human	Mouse		
20	Agrin exon 32 / AGRN	1. 247	1. 257	does not exist	false
		2. 251	2. 265	does not exist	false
		3. 365			false
		4. 445			false
		5. 620			false
		6. 642			false
		7. 734			false
		8. 885			false
21	Type II activin receptor / ACVR2	-------	-------		
22	GABA gamma2 / GABRG2	1. 31	1. 31	exists	true
23	c-src exon N / CSK or SRC	1. 66	1. 57	exists	true
		2. 289	2. 622	does not exist	false
24	Agrin exon 28 / AGRN	-------	1. 875	does not exist	false
25	FE65 / APBB1	-------	-------		

Discussion

We used the following criterion to distinguish the functional from the nonfunctional UGCAUG sites: if a site is conserved, it is functional. This criterion is based on the assumption that regulation of alternative splicing of orthologous mouse and human genes is similar. We do not assume, however, that all the non-conserved motifs are not functional.

Conserved hexanucleitides were flanked by conserved sequences on both sides. Some were close to the exon (Table). Thus, it is likely that the UGCAUG signal is a part of a cis-regulatory system. One such example is provided by the mouse *c-src* gene. The so-called downstream control sequence in the intron downstream of the cassette exon includes UGCAUG and two other elements. The KSRP and FBP proteins were shown to bind to UGCAUG, PTB or nPTB bind to pyrimidine tract, KSRP and FBP bind to the third regulatory element (Ladd, Cooper, 2002).

Acknowledgements

The work was supported by grants HHMI (55000309), LICR (CRDF RBO-1268), RFBR (04-04-49440), the Fund for Support of the Russian Science and the program “Origin and Evolution of the Biosphere” of RAS. We are grateful to A.A. Mironov, R. Nurtdinov and I. Kosmodemiansky for helpful discussions.

References

Brudno M., Gelfand M.S., Spengler S., Zorn M., Dubchak I., Conboy J.G. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing // Nucl. Acids Res. 2001. V. 29. P. 2338–2348.

Ladd A.N., Cooper T.A. Finding signals that regulate alternative splicing in the post-genomic era // Genome Biol. 2002. 3(II), reviews008.1-0008.16.

Sorec R., Ast G. Intronic sequences flanking alternatively spliced exons are conserved between human and mouse // Genome Res. 2003. V. 13. P. 1631–1637.

Waterson R.H., Lindblad-Toh K., Birney E., Rogers J., Abril J.F., Agarwal P., Agarwala R., Ainscough R., Alexandersson M., An P. *et al*. Initial sequences and comparative analysis of the mouse genome // Nature. 2002. V. 420. P. 520–562.

INFORMATIONAL ASPECTS OF THE LATENT TRIPLET PERIODICITY ANALYSIS

Frenkel F.E.*, Chaley M.B., Korotkov E.V., Skryabin K.G.

Centre "Bioengineering" RAS, Moscow, Russia
* Corresponding author: e-mail: felix@biengi.ac.ru

Keywords: *triplet periodicity, classification, biological database*

Summary

Motivation: Modern biological data having large size at the same time requires intensive statistical and structural analysis. While such intensive calculations are well taken by computational clusters statistical analysis of a homogenous objects is implemented on relational databases in a natural way.

Results: Some informational aspects on systematisation of search and analysis of triplet periodicity in the nucleotide sequences are present. Database on triplet periodicity found in the genes from KEGG (Kyoto Encyclopaedia of Genes and Genomes) has been created. A general scheme of data processing, structure of developed database and sample results are also shown.

Introduction

At present time analysis of genetic sequences accomplished by mathematical methods leads to huge body of materials that could not be handled without making specialised informational systems for storage and processing of data obtained. The present work shows creation of such an instrument to help experts-biologists in understanding of gene's thin structure and evolution. An object of the investigation is the triplet periodicity (periodicity having period of three symbols) in the gene sequences. Such a periodicity is typical for DNA regions encoding proteins. The created database provides correlation analysis between the classes of triplet periodicity in the genes and the functions of their encoded proteins along with organisms' taxonomy.

Model

As data source for analysis we have used Kyoto Encyclopaedia of Genes and Genomes (KEGG) [4]. An overall computing process could be schematically shown as three stages of data stream processing (see Fig. 1):

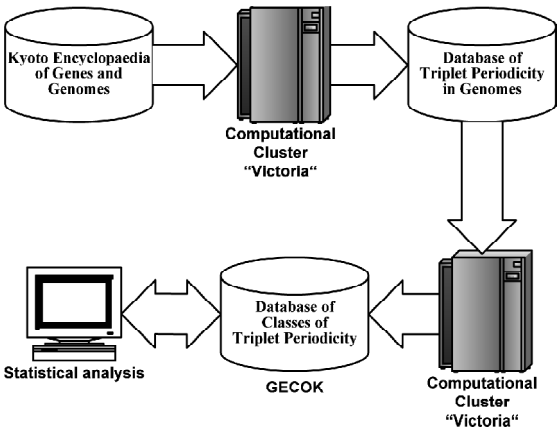


Fig. 1. Overall computing process.

1. Search for the triplet periodicity in the nucleotide sequences of genes in the source database (KEGG).
2. Classification of found triplet periodicity.
3. An analysis of the obtained triplet periodicity classes with respect to functional protein classes and taxonomy of organisms).

The software created during the first two stages has been implemented in Fortran language using MPI standard's library of inter-process communications. This method of development has allowed use as much as possible the computational resources of Victoria cluster where the

programs have being ran. This cluster is situated at the Centre "Bioengineering" of RAS, Moscow. It consists of the 32 Pentium IV (2,4 GHz) nodes with 1 Gb RAM and 40 Gb disk drive (for temporary files storage). Head of the cluster is dual processor (2 x Intel Xeon 2.4 GHz, 1 Gb RAM, 72 Gb RAID-1(2 x Ultra320 SCSI Wide)) server that is connected with nodes by Gigabit Ethernet network. All the nodes are additionally connected by their own 100 MBit Ethernet network (see Fig. 2). There is RedHat Linux 7.3 OS on all cluster nodes and the server. The server also has GNU GCC compilers toolkit installed with MPICH library [5] for inter-process communications.

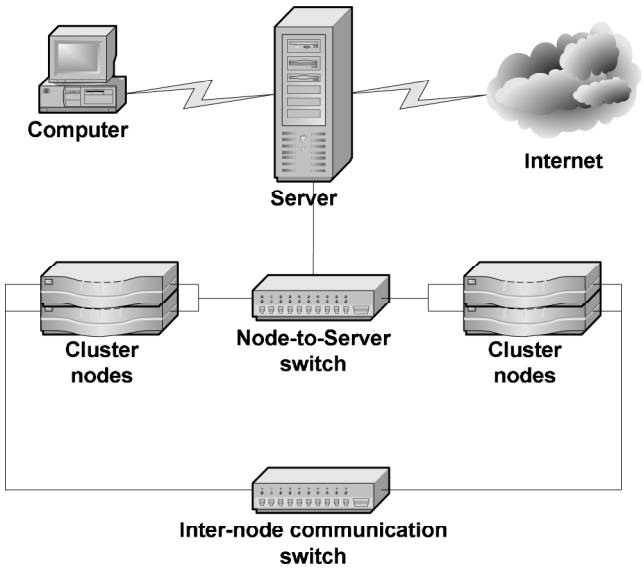


Fig. 2. Cluster network architecture.

Results and Discussion

General database structure has five key entities: found regions of triplet periodicity, classes of these periodicity regions, genes, organisms, and taxonomic divisions where periodicity has been found.

Database created under Firebird RDBMS [2] has been filled by software written in Java language. To access the database we use IB Expert [3] manager providing full access to database contents and their visualisation.

Table below contains sample results obtained from the created database. Detailed description of the applied algorithms and analysis of data biological meaning could be found in [1].

Table. Share of periodic genes, and number of periodicities classes in each KEGG division

Kegg division	A number of available genes	A number of periodic genes	Periodic/ Available, %	Species
Protists	16787	12620	75	
Animals	66404	53868	81	
Plants	26960	20809	77	
Fungi	14042	10150	72	
Archaea	39185	29130	74	
Bacteria	253051	194499	77	
Totally	416429	321085	77	

References

1. Chaley M.B., Frenkel F.E., Korotkov E.V., Skryabin K.G. Relationships between General classification of genes, latent triplet periodicity and The universal phylogenetic tree // BGRS'2004.
2. Firebird Project Homepage. http://firebird.sourceforge.net/.
3. IB Expert Homepage. http://www.ibexpert.com/.
4. Kyoto Encyclopedia of Genes and Genomes. http://www.genome.ad.jp/kegg/.
5. MPICH Project Homepage. http://www.mcs.anl.gov/mpi/mpich/.

A PRACTICAL METHOD FOR MAXIMUM EXACT MATCHES IN LARGE GENOMES

***Fursov M.Yu.*[1], *Baksheyev D.G.*[*1], *Rodionov K.V.*[1], *Golubitskii A.A.*[1], *Saraev D.V.*[2], *Denisov S.I.*[2], *Blinov V.M.*[2]**

[1] Novosibirsk Center of Information Technologies "UniPro", Novosibirsk, Russia; [2] State Research Center of Virology and Biotechnology "VECTOR", Koltsovo, Russia
* Corresponding author: e-mail: fmike@bits.unipro.ru; bd@bits.unipro.ru

Keywords: *whole genome alignment, dot matrix*

Resume

Motivation: The search for exact matches of matches of substrings in pairs of large genomic sequence is inbuilt and performed by many genomic analysis programs. Theoretically, the problem has a linear time solution based on building suffix tree or suffix array. For large sequences, the size of this indexing structure often exceeds the available computer memory, thereby making a desktop analysis of large genomes extremely time-consuming. In most cases, the use of these fast algorithms can be appreciably extended with incomplete sequence indexing.

Results: We describe a new fast and memory efficient indexing method for searching maximum exact matches in large genomes. The method is based on gapped suffix indexing and requires several times less memory and time than the traditional methods. With the gapped indexing, an interactive analysis of whole large genomes becomes executable on a desktop computer. The method serves as the basis for DPview, the system for searching, visualization and analysis of similarities in whole large genomes.

Availability: http://bits.unipro.ru/eng/dpview.html

Introduction

Genomic sequence numerous is one of the most important methods for genome analysis and it provides applications in genome assembling, searching functional elements, establishing evolutionary relationships, and defining genome structure. With more whole eukaryotic genomes becoming available, special tools are required for comparing sequences of hundreds megabases and larger. In the past years, many tools were developed for pairwise and multiple sequence alignments without limiting the length of the sequences to be aligned (Schwartz *et al*., 2000; Bray *et al*., 2003; Brudno *et al*., 2003; Kurtz *et al*., 2004). However, the application of tools to analysis of large genomic fragments is made difficult by the need for striking a balance between the sensitivity and vapidity of the algorithms. It has become a tradition to reachthis balance by using the anchor-based approach (Batzoglou *et al*., 2000).

As to pair alignment, this approach first calculates possible anchors, large non-overlapping substrings of two strings that are very similar. The anchors form the basis for alignment, so one has only to align the sequences using slower but more precise methods of local alignment. Traditionally, the very first step of anchor search is the search for all the maximum exact matches (MEMs) of length not less than a fixed w (Bray *et al*., 2003; Kurtz *et al*., 2004). Efficient implementation of this operation contributes to the total performance of the global alignment algorithms.

Theoretically, the task of searching exact matching substrings between two strings has the linear time solution (Gusfield, 1997) based on building an index structure or index, such as suffix tree (Kurtz *et al*., 2004), suffix array (Manber, Myers, 1993), hash-table (Ning *et al*., 2001) or trec-based structure (Brudno *et al*., 2003). In all the cases, the size of the index is several times (4 and more) that of the sequence itself.

Another traditional method for searching maximum exact matches of length at least w is the generate-

and-test approach first proposed in (Dumas, Ninio, 1982). It relaxes memory requirements at the expense of extra computations. In that method, all *k*-mers, the exact matches of certain fixed length $k < w$, are first generated and then tested if each can be extended up to the length of not less than *w*. This extension is accomplished by pairwise character comparison around the *k*-mers and, therefore, takes the time proportional not only to the number of the matches found but also to their length.

What we ultimately have are the algorithms based on full suffix indexing, optimal in time but requiring large memory, on the one hand, and not very fast but much more memory-efficient algorithms for seed extension, on the other hand. We offer an approach that combines the advantages of both traditional kinds of algorithms.

Method and Algorithm

To solve the problem of searching maximum exact matches in a pair of two sequences we suggest to use a modified version of indexing structure, gapped suffix. In contrast to the creation of a total suffix tree (Kurtz *et al*., 2004), gapped indexing of suffixes allows us to gain time and memory at the subsequent the step with little impact on the performance of subsequent comparisons.

The problem we solve is to find all maximum exact matches of length $l \geq w$ for given strings $A=A[1..|A|]$ and $B=B[1..|B|]$ and an integer $w > 0$. To accomplish this, we will build a suffix indexing structure *S* (suffix tree or suffix array, shortly "index") only for sequence *A*.

The whole index is created as follows: let us denote by $A_i=A[i..|A|]$ the suffix of string *A*. After all the suffixes of *A* are indexed in a structure *S*, we will move sequentially along string *B* and use *S* to search for such suffixes B_j that have prefixes matching the suffixes of string *A* in the first *w* characters or more. Assume that we have found a result $(i_0; j_0; l)$, that is such j_0 that $B[j_0 - 1] \neq A[i_0 - 1]$ and $B[j_0..|B|]$ has a common prefix of $l \geq w$ with $A[i_0..|A|]$. Then, $l - w + 1$ following single-character shifts along string *B* will result in the matches of suffixes B_j and A_i in the first $l - (j - j_0)$ characters, thereby giving the matches that are the parts of this match but not the maximal. To avoid such repetitive comparisons, we first save the matches in a buffer structure *Q*.

The idea of our method is that indexing sequence *A* with a step *s*, $0 < s \leq w$, and then expanding the matches found to their maximum is sufficient to apply the approach described above. In this case, by checking the suffixes of string *B* in order and finding the ones that match the suffixes of *A* included in *S* in at least $v = w - (s - 1)$ first characters we will not miss any single match of the length *w*. If the match found is not a part of an already obtained match, then, after extending it to the left by $k \leq s$ characters so that the maximum match would be obtained, it should be saved as a result if its length is $k + v \geq w$.

Thereby we reduce the size of the indexing structure *S* and the time required for its creation. The need for additional pairwise comparisons at the expansion step is counterweighted by the fact that *S* stores the information about only one of *s* suffixes, and thus is of a much smaller size issignificantly smaller size. Repetitive expansion of earlier matches is suppressed in our algorithm by a filtering of the results.

Results and Discussion

The described method of searching maximum exact matches requires significantly less memory, it reduces the time needed for creating a suffix structure *S* and the search time in the case of a suffix array. For instance, building a gapped suffix array for a string of 100 Mb for w=50 and s=25 is equivalent to building a full suffix array for a sequence of length 100/s = 4 Mb. By increasing gap *s* we can shrink the memory used to store the suffix even more, but it also will increase the number of seed matches found of length $v=w(s\ 1)$ and, in consequence, will increase the resources needed for filtering operation, the size of buffer *Q*. In order to find the optimal number of gaps one should take into consideration the memory available, the size of the sequence alphabet, expected number of repeats in the string to be indexed, and the overall complexity of the algorithms to build and search the index.

Table shows results of comparison of our approach, implemented in UniPro DPview software (freely

available at http://bits.unipro.ru), and MUMmer (Kurtz *et al.*, 2004), the most well known system able to compute maximum exact matches in large genomes. All tests were run on Pentium III 650 MHz with 1Gb RAM. Testing results show in detail practical applicability and effectiveness of the method suggested by our group. It should be noticed that DPview is a successor of MegaGene software (Blinov *et al.,* 1999, 2001) and outperforms it by several orders on the task of finding MEMs. A whole human vs. chimpanzee genome alignment obtained with DPview is shown in Figure.

Table. MUMmer vs. DPview — searching MEMs of length *l* e" *w*

Sequences	Size, Mbp^2	*w*	MUMmer		DPview	
			Sec	Mb	Sec	Mb
E.coli.NC_002655 × self	5.47 ×5.47	20	23	94	7	11
		30	23	94	5	9
Human.chrY.NT_011875 × self	10×10	30	71	169	24	16
		45	110	400	15	14
Human.chrX.NT_011757 × self	24×24	35	134	400	49	34
		45	110	400	32	31
Human.chrX.NT_011651 × self	31 × 31	40	-[a)]	-[a)]	54	42
Human.chrY.NT_011875 × Mouse.chrX	10 × 150	50	310	308	60	198
Mouse.chrX × Rat.chrX	150 × 160	90	-[b)]	-[b)]	327	340

[a)] Despite sufficient RAM, MUMmer failed because of memory usage peaks at the tree creation step.
[b)] Theoretically, insufficient memory for MUMmer.

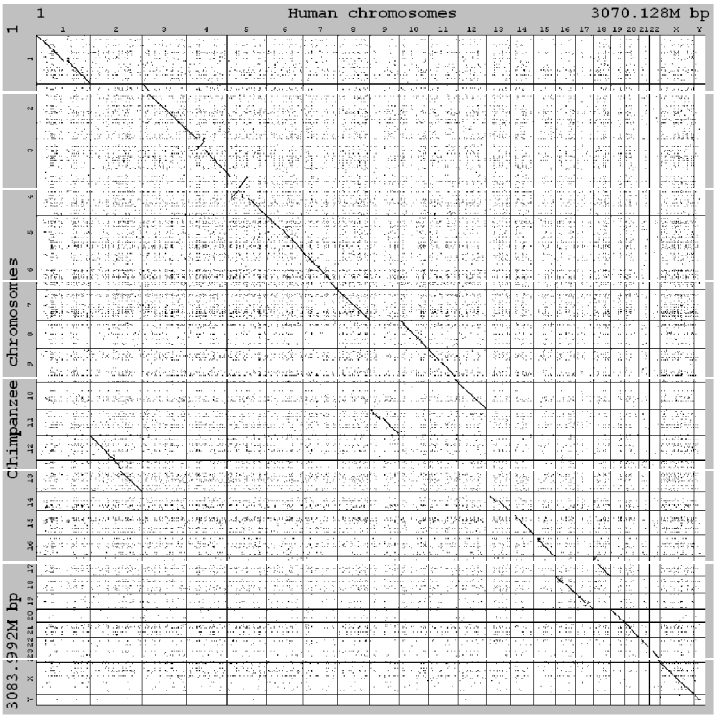


Fig. Alignment of the human genome vs. the chimpanzee genome.
Maximum exact matches of length 350 Mb and more between human (NCBI Build 34) and chimpanzee (UCSC Build 1) genomes (both DNA strands). The computation took about 20 hrs on Pentium III 650 MHz with 1Gb RAM.

Acknowledgments

The authors are very grateful to Ivan S. Golosov (UniPro) for financial support of the research and to Lev Kisselev, Full Member of the Russian Academy of Science, for the equipment provided for testing and debugging the program. Chimpanzee Genome Sequencing Consortium is acknowledged for making the genome publicly available.

References

Batzoglou S., Pachter L., Mesirov J.P., Berger B., Lander E.S. Human and mouse gene structure: Comparative analysis and application to exon prediction // Genome Res. 2000. V. 10. P. 950–958.

Blinov V.M., Denisov S.I., Saraev D.V., Shvetsov D.V., Uvarov D.L., Oparina N.Yu., Sandakhchiev L.S., Kisselev L.L. Structural organization of the human genome: distribution of nucleotides, Alu repeats, and exons in chromosomes 21 and 22 // Mol. Biol. 2001. V. 35(6). P. 1–7. (in Russian).

Blinov V.M., Resenchuk S.M., Denisov S.I., Chirikova G.B., Uvarov D.L., McCorkle S., Anderson C. MegaGene: a New Computer Technology for Analyzing Complete Viral Genomes // XI-th International Congress of Virology, Sydney. 1999.

Bray N., Dubchak I., Pachter L. AVID: A global alignment program // Genome Res. 2003. V. 13. P. 97–102.

Brudno M., Chapman M., Gottgens B., Batzoglou S., Morgenstern B. Fast and sensitive multiple alignment of large genomic sequences // BMC Bioinformatics. 2003. V. 4. 66.

Dumas J.P., Ninio J. Efficient algorithms for folding and comparing nucleic acid sequences // Nucleic Acids Res. 1982. V. 10. P. 197–206.

Gusfield D. Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology. Cambridge University Press, UK. 1997.

Kurtz S., Phillippy A., Delcher A.L., Smoot M., Shumway M., Antonescu C., Salzberg S.L. Versatile and open software for comparing large genomes // Genome Biology. 2004. V. 5(2). R12.

Manber U., Myers G. Suffix Arrays: A new method for on-line string searches // SIAM J. of Computing. 1993. V. 22(5). P. 935–948.

Ning Z., Cox A.J., Mullikin J.C. SSAHA: a fast search method for large DNA databases // Genome Res. 2001.V. 11(10). P. 1725–1729.

Schwartz S., Zhang Z., Frazer K.A., Smit A., Riemer C., Bouck J., Gibbs R., Hardison R., Miller W. PipMaker A web server for aligning two genomic DNA sequences // Genome Res. 2000. V. 10. P. 577–586.

DISTRIBUTION OF THE SF-1 BINDING SITES PREDICTED BY THE SITEGA METHOD IN THE GENOMIC SEQUENCES AND THEIR EXPERIMENTAL VERIFICATION

Ignatieva E.V. *, Levitsky V.G., Vasiliev G.V., Klimova N.V., Busygina, T.V., Merkulova T.I.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: eignat@bionet.nsc.ru

Keywords: *SF-1, transcription factor binding site recognition, steroidogenic genes regulation*

Summary

Motivation: The SF-1 transcription factor is the key regulator of the steroidogenic genes expression. Identification of new target genes for the factor in vertebrate the genome is required for a better understanding of the regulatory mechanisms of the biosynthesis of steroid hormones and of the entire reproductive system. It is expedient to minimize the expenditures for large-scale experimental tests of the candidate genes identified by computer methods. For this purpose, preliminary tests of the efficiency and accuracy of the recognition methods for the SF-1 sites are required.

Results: Analysis of the predicted SF-1 binding sites distribution in the genomic sequences demonstrated that the SiteGA method models well a natural situation. The content of the SF-1 sites in the promoter regions of the genes for the steroidogenic system was found to be high as compared with that of genes belonging to other functional groups. The most contrasting group (the cell cycle genes) contains no potential SF-1 sites in all the analyzed regions (the 5'flanking regions, exons, introns). The experimental tests demonstrated that 83 % of the sites recognized by the SiteGA method, indeed, interact with the SF-1 factor.

Introduction

The transcription factor SF-1 plays a key role in the regulation of the steroidogenic genes and is required for the normal development of the pituitary-adrenal and gonadal axis (Busygina *et al*., 2003; Val *et al*., 2003). SF-1 is a member of the nuclear receptor superfamily. In contrast to most of its members, interacting with DNA as a homodimer or a heterodimer, SF-1 activates gene expression by binding to DNA in a monomeric form (Val *et al*., 2003). There have been, so far, no computer-assisted methods concerned with the recognition of the SF-1 binding sites. Here, we report the results of potential SF-1 binding sites identification by using the SiteGA recognition method (Levitsky *et al*., 2004, this issue). The testing of the predictions involved two steps. At the first step, we estimated the content of the predicted SF-1 sites in the genomic sequences with contrasting features, for example, in the promoter regions of the steroidogenic genes (which are abundant in SF-1 sites) and also the promoter regions of genes belonging to other functional groups. We also estimated the density of the predicted SF-1 sites in the non-promoter sequences (exons and introns), which have a smaller content of transcription regulatory elements than promoter regions. At the second step, the capability of the predicted SF-1 sites to interact with this protein was experimentally tested.

Methods and Algorithms

Samples of the SF-1 binding sites, and the regulatory regions, exons, and introns of genes, beloning to different functional groups. To develop the recognition method, we used a sample of the nucleotide sequences of 54 experimentally identified SF-1 binding sites from the TRRD (Kolchanov *et al*., 2002). The samples of the promoter (the C_CYC, INTERF, LipMet, END_P,

MACROPH, LIVER, and ERITHR groups described in the legend to Fig. 3) were set up on the basis of the data on the position of the transcription start stored in the TRRD (Kolchanov *et al*., 2002). The STER sample contained the promoter regions of steroidogenic genes whose transcriptional regulation by SF-1 has not been, as yet, experimentally studied. The major source for establishing the STER sample were the EMBL database and data on the literature referring a gene to the steroidogenic system. The samples of exons and introns were extracted from the EMBL database.

The SiteGA method for the recognition of the SF-1 site. To build functions for the SF-1 site recognition, the entire analysed 93 bp region was divided into 3 parts: the left flank, the center, and the right flank (36 bp long each). To identify the contextual DNA features determining the presence of the SF-1 sites, we applied a model (Levitsky, Katokhin, 2003) based on the correlation between the mutual disposition of sites of particular dinucleotides with the local parts of the site region into which each of the site regions was subdivided (Levitsky *et al*., 2004, this issue). Given this subdivision for each of the 3 regions, we determined the optimum significance levels (thresholds) for the recognition functions of the 3 areas (Table 1).

Table 1. The optimum significance levels for the recognition function in each of the regions of the SF-1 site

Threshold set number	Significance levels for		
	central area	right flank	left flank
I	0.95	0.90	0.95
II	0.95	0.80	0.70
III	0.80	0.80	0.70

Experimental verification of the potential SF-1 sites. SF1 binding to 32 bp labeled double-stranded oligonucleotides corresponding to the predicted sites was assayed by the EMSA A nuclear extract from testicular cells from 14 day-old male Wistar rats served as the SF-1 source. The DNA/protein complexes truly incorporate SF-1, as evidenced by the disappearance of the corresponding band after the addition of antibodies against SF-1 (The Upstate Biotechnology).

Implementation and Results

***Distribution of the binding sites for SF-1 that were found experimentally relative to the transcription start* (Fig.1).** According to the TRRD data, the SF-1 sites are located in rather long regions (from -4820 to +155). However, the majority of SF-1 sites are located at the 5'-flanking region of the gene adjacent to the transcription start; 83 % (45 of 54) lie in the region between –300 and -1, 50 % (27 of 54) occupy the region between –150 and –50.

Estimation of the content of the potential SF-1 sites in the promoter regions (between –300 and -1) of genes, the members of different functional groups. At the chosen significance levels (Table 1), the number of the SF-1 sites predicted by the SiteGA method in the regions of the steroidogenic genes (the STER sample) was 2.0, 1.7 and 1.5 per 1,000 bp (Fig. 2). The content of the potential SF-1 sites was smaller in the promoter regions of the other gene groups. The genes of ERYTHR sample contained 2–3 times less SF-1 sites than the STER sample. The content of the SF-1 sites was still smaller in the LIVER, LipMet, MACROPH, and INTERF samples. There were no potential SF-1 sites when all the three sets of significance levels were applied to the C_CYC sample containing 75 promoter regions of the cell cycle genes. Depending on the chosen set of significance levels, the STER sample had an estimate exceeding the one for the combined promoter sample (which included the remaining 293 promoters) by 5.1–10.9 times. The greatest difference (10.9-fold) was observed for the set of significance levels II.

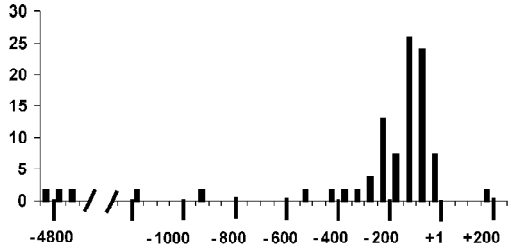


Fig. 1. The position of the SF-1 sites stored in the TRRD database relative to the transcription start. The distance from the transcription start is plotted along the abscissa. The number of sites (as percentage of the total site number) is plotted along the ordinate.

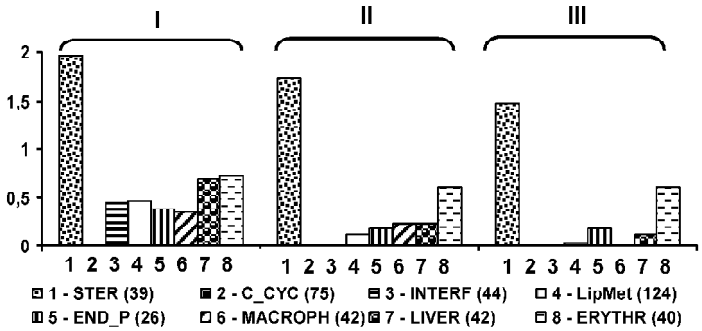


Fig. 2. An estimation of the potential SF-1 sites content in the promoter (300/-1) regions of genes, the members of different functional groups. STER – the steroidogenic genes; C_CYC - the cell cycle genes; INTERF – the interferon regulated genes; LipMet – the genes controlling lipid metabolism; END_P – the genes expressed in the endocrine pancreas; MACROPH – the genes expressed in macrophages; LIVER – the genes, expressed in liver; ERYTHR – the genes expressed in erythroid cells; The number of genes in each group is in parentheses. I, II, III designates the sets of the significance levels used for the recognition. The number of the SF-1 sites per 1,000 bp is along the ordinate.

Estimation of the content of the potential SF-1 sites in the 5'-flanking regions, exons and introns by the SiteGA method, using the set of significance levels II. The steroidogenic genes, for which SF-1 is the key regulator, contained the highest number of potential SF-1 sites in the region between –300 and –1 (Fig. 3). The lipid metabolism genes showed the densest putative SF-1 sites in the region between –900 and -600, however, even in this region, their number was 3 times smaller than in the region between –300 and -1 of the steroidogenic genes (Fig. 3). The content of the SF-1 sites in exons and introns of the steroidogenic and lipid metabolism genes was, on average, four times smaller than in the region between –300 and -1 of the steroidogenic genes (Fig. 3). The comparatively high prediction level for the SF-1 sites in the genes of lipid metabolism, as well as for the genes expressed in liver, is probably due to the associated identification of a number of LRH1 site by SiteGA. LRH1 is a close homolog of SF-1 and participates in the regulation of the above genes (Val *et al*., 2003). The extremely low level of potential SF-1 sites (0.047 and 0.132 per 1,000 bp) was observed for non-coding exons and introns of the cell cycle genes. It is noteworthy that the cell cycle genes contain no SF-1 sites both along the entire analysed 5'regions (between –900 and -1) and in coding exons.

The new potential SF-1 sites identified by SiteGA and their experimental verification. We analysed a sample of 5'-flanking sequences of 33 steroidogenic genes for which an experimental search of the SF-1 sites has, as yet, not been performed. When using the set with significance level II, 12 potential sites (Table 2) were identified in the promoter regions of 8 genes (mouse 3betaHSDI and Ad4BP/SF-1, ovine, macaque, and bovine StARs, porcine CYP17 and LHbeta, rat HSD17BI). In addition, potential sites were searched for in non-coding exons and introns of several genes. As a result of analysis of exons and introns, 1 potential site was found in the non-

coding exon of the human HSD17BII and 7 sites were identified in introns of the mouse Slp, and bovine StAR. Gel shift experiments using specific antibodies demonstrated the high accuracy of the SiteGA method. In general, the capability of the predicted SF-1 sites to interact with protein was confirmed for 83 % (15 of 18) of the computionally predicted sites. For the sites, localized in the promoter regions, the proportion of the experimentally supported sites was higher still, 90 %.

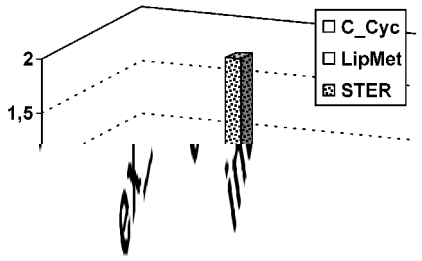


Fig. 3. The number of SF-1 sites predicted using the SiteGA method in the 5'flanking regions, exons, and introns of the cell cycle (C_CYC), the lipid metabolism (LipMet), and the steroidogenic (STER) genes. The number of SF-1 sites per 1,000 bp predicted using the set of significance levels II is plotted along the ordinate.

Table 2. The results of the recognition of the SF-1 sites in the steroidogenic genes by the SiteGA method using the set of significance level II and its experimental support

Gene region*	Gene number	Site number		
		recognized	tested experimentally	experimentally supported
Promoter (13129)	33	12	10	9
Non-coding exons (5109)	10	1	1	1
Introns (16553)	13	7	7	5

* Total length (bp) is in parentheses.

Discussion

The SiteGA method recognizes the SF-1 sites taking into account the contextual properties of a region with a total length of 93 bp, covering both the conserved and the flanking regions (Levitsky, 2004, this issue). The subset of the optimum significance levels considered separately for each of the 3 regions (the central, the right, and the left flanks) allowed us to considerably increase recognition accuracy. Analysis of the SF-1 site in the genomic sequences predicted by the SiteGA method demonstrated that it models well natural situations. The steroidogenic genes for which SF-1 is the key regulator (Busygina *et al.*, 2003) show the highest density in the SF-1 sites in the regulatory regions (Fig. 2). It was found that the region between –300 and -1 relative to the transcription start site is most abundant in the SF-1 sites (Fig. 3). These results are in good agreement with those for the localization of real sites (Fig. 1). Using the SiteGA method, it was also shown that the distribution of the potential SF-1 sites in the 5'flanking regions of the genes, in coding and non-coding exons, and in introns is specific to the gene functional groups (Fig. 3). Experimental tests of the SF-1 sites predicted by the SiteGA method supported its high efficiency (83 % of the SF-1 sites were correctly predicted). The highest proportion of false predictions was in introns. This was probably because the training sample for the SF-1 sites was taken from the promoter regions. The experimentally supported sites were located in the regulatory regions, exons and introns of 10 genes relevant to the steroidogenic system. Our data indicate that these genes are

very likely the targets for the SF-1 factor. It should be also noted that 5 of the 15 experimentally supported SF-1 sites were located in introns, 1 site was in non-coding exon. The localizations contribute to our understanding of the role of the internal regions in gene regulation.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (No. 03-04-48506-a, 03-07-90181-в 03-04-48469-a), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), SB RAS Integration Project No. 119, RAS Presidium Program "Molecular and Cellular Biology" (project No. 10.4). The authors are grateful to Ananko E.A., Podkolodnaya O.A., Turnaev I.I., Voronich E.S., Proskura A.L. Pozdnyakov M.A., Lokhova I.V. for technical support and to Kolchanov N.A., Osadchuk A.V. for helpful discussions.

References

Busygina T.V., Ignatieva E.V., Osadchuk A.V. Consensus sequence of transcription factor SF-1 binding site and putative binding site in the 5;-Flanking regions of genes encoding mouse steroidogenic enzymes 3betaHSDI and Cyp17 // Biochemistry (Mosc.). 2003. V. 68. P. 377–384.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acid Res. 2002. V. 30. P. 312–317.

Levitsky V.G., Ignatieva E. V., Busygina T.V., Merkulova T.I. Analysis of the context features of SF-1 binding site and development of a criterion for SF-1 regulated gene recognition by the SiteGA method // This issue, 2004.

Levitsky V.G., Katokhin A.V. Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis // In Silico Biol. 2003. V. 3. 0008. http://www.bioinfo.de/isb/2003/03/0008/

Val P., Lefrancois-Martinez A.M., Veyssiere G., Martinez A. SF-1 a key player in the development and differentiation of steroidogenic tissues // Nuclear Receptor. 2003. V. 1. P. 8–45.

COMPARISON OF THE RESULTS OF SEARCH FOR THE SF-1 BINDING SITES IN THE PROMOTER REGIONS OF THE STEROIDOGENIC GENES, USING THE SITEGA AND SITECON METHODS

Ignatieva E.V.*, Oshchepkov D.Yu., Levitsky V.G., Vasiliev G.V., Klimova N.V., Busygina T.V., Merkulova T.I.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: eignat@bionet.nsc.ru

Keywords: *binding site recognition, SF-1, conformational and physicochemical DNA properties, discriminant analysis*

Summary

Motivation: Development of accurate prediction methods for the transcription factor binding sites (TFBSs) is important in investigation of the regulatory regions of the eukaryotic genes. In relation to this topic, of great interest appears comparative analysis of recognition methods, based on different rationales with respect to the binding sites of the same transcription factor.

Results: We compared two new TFBSs recognition methods SiteGA and SITECON, providing search for the SF-1 sites in the 5'-regions of the steroidogenic genes as an example. 22 new potential SF-1 binding sites were identified in 17 of the 33 genes. Experimental tests demonstrated the high efficiency of the recognition methods: 19 of the 20 experimentally checked sites are capable of binding to SF-1 *in vitro*.

Introduction

Development of computer methods for recognizing the TFBS is a promising approach to a better understanding the regulatory DNA code. Currently, the development of such methods is based on different approaches and algorithms, which are used, as a rule, to search for TFBS of a particular transcription factor. This makes very difficult comparative analysis of the performance of the various methods. Here, we present the results for SF-1 (steroidogenic factor 1) prediction in the promoter regions of the same genes of the steroidogenic system, using two new SiteGA and SITECON methods. The two methods have been developed on the basis of analysis of the SF-1 sites from the TRRD (Kolchanov *et al*., 2002). To estimate the recognition quality, we tested the capability of the predicted sites to interact with the SF-1 protein. The results supported the high recognition accuracy of the two methods.

Methods and Algorithms

Samples of nucleotide sequences of the SF-1 binding sites and the 5'-flanking regions. To develop the recognition methods, we used a sample of the nucleotide sequences of 54 experimentally identified SF-1 binding sites from the TRRD (Kolchanov *et al*., 2002). Search for the new SF-1 binding sites was done in the 5'-flanking regions of 33 steroidogenic genes without experimentally supported SF-1 sites in their regulatory regions. A part of the 5'regions of the genes was included in the sample on the basis of the TRRD data on the positions of the transcription start; however, the main sources for setting up the sample were the data on the positions of the transcription start from the EMBL nucleotide sequence database.

The SiteGA method for SF-1 site recognition. To build the function for SF-1 site recognition, the entire analysed 93 bp region was subdivided into 3 parts: the left flank, the center, and the right flank (36 bp each). To reveal the context features of DNA determining the presence of the SF-1 sites we used the model (Levitsky, Katokhin, 2003) based on the correlations between the mutual

disposition of the particular nucleotides within the local site region resulting from subdivision of each of 3 site regions (Levitsky *et al*., this issue). The optimum significance level for the recognition functions of the 3 regions was calculated: 0.95 (the center), 0.8 (the right flank), 0.7 (the left flank).

The SITECON method for SF-1 recognition. SITECON is a method for recognizing sites based on analysis of their conserved physicochemical and conformational properties (Oshchepkov *et al*., 2004). As the recognition threshold, the SITECON method employs the level of necessary conformational similarity (Oshchepkov *et al*., 2004), which was 94 % for SF-1. The recognition quality for type I errors was checked with the jack-knife method, which removes 1 sequence in series from the training sample. Recognition of binding sites in the negative sequence with a length of 500,000 bp was the control for the type II error. The negative sequence was generated by random shuffling of nucleotides in the initial site sequences; thus, the nucleotide content of both the positive and negative samples was identical, and the recognition was carried out on both strands. Estimates for type I and II errors at these levels of conformational similarities (92 %, 93 %, 94 % and 95 %) are given in Table 1.

Table 1. Errors for SF-1 binding site recognition by the SITECON method, calculated for different levels of conformational similarity

	92.00%	93.00%	94.00%	95.0%
Type I error	0.30	0.39	0.56	0.70
Type II error	7.31E-04 (1/1368)	5.22E-04 (1/1915)	2.23E-04 (1/4484)	6.97E-05 (1/14347)

Experimental confirmation of the potential SF-1 sites. SF1 binding to 32 bp labeled double-stranded oligonucleotides corresponding to the predicted sites was assayed by the EMSA. A nuclear extract from testicular cells of 14 day-old male Wistar rats served as the SF-1 source.

Implementation and Results

Identification of new SF-1 potential binding sites in the promoter region of the genes by the SiteGA and SITECON methods. Analysis of the promoter regions in 33 genes revealed 22 new SF-1 sites (Table 2); of these, 5 were recognized by both methods, 7 only by SiteGA, and 10 only by SITECON.

Experimental checking of the potential SF-1 sites by the electrophoretic mobility shift assay (EMSA). Twenty of the 22 recognized sites (Table 2) were checked experimentally. Experimental support was obtained for 90 % of the sites (9 of 10) predicted by SiteGA and for 100 % of those predicted by SITECON. Illustrative results provided by the EMSA are shown in Fig.

Discussion

Experimental analysis demonstrated the high accuracy of the two methods for the recognition of the SF-1 site. The capacity to interact with SF-1 was supported for 90 % of the sites predicted by SiteGA and for 100 % of those predicted by SITECON. However, the site sets predicted by each method alone overlapped only partly. In all, 5 sites were identified by both methods. Thus, the proportion of "common sites" in the group recognized by SiteGA was 42 % (5 of 12), and it was 33 % (5 of 15) in the group, recognized by SITECON. Partial overlapping of the prediction results is easily explained when one takes into consideration the following: (i) To minimize the number of false predicted sites (Type II error) for both methods, we chose a threshold with type I prediction error exceeding 50 %; (ii) in the model of the site, the applied SiteGA and SITECON are different, in principle, in the course of SF-1 site recognition. The model used for the SiteGA method includes a description of the correlations (of which ~400 are significant) between the dinucleotide frequencies within the local site region into which the 3 site regions were subdivided (Levitsky *et al*., 2004, this issue). The model of the SF-1 site on which recognition by SITECON is based contains data on a high number of positions with highly conserved

conformational and physicochemical properties of the DNA double helix. Although the results overlapped only partly, virtually all the sites recognized by SiteGA and SITECON were supported by experimental data. With this in mind, it appears expedient to implement at the same time recognition of the SF-1 sites in sequences under study by the two methods. The SF-1 binding sites predicted and supported by experimental data are located in the regulatory regions of 17 vertebrate genes, whose SF-1 regulation has not yet been studied experimentally. Our results indicate that the expression of the 17 genes is most likely regulated with the involvement of the SF-1 factor.

Table 2. Potential SF-1 binding sites recognized in the promoter regions of the steroidogenic genes by the SiteGA and SITECON methods and experimental support

	Gene	Position*	Forward or reverse	Recognized by		Experimental support
				SiteGA	SITECON	
1	*Cyp17*(Mouse)	-283	<-	-	+	+
2	*Cyp17* (Mouse)	-49	->	-	+	+
3	*HSD3b* (Mouse)	–113	<-	+	+	+
4	*Cyp11B1* (Guinea pig)	-126	<-	-	+	+
5	*Ad4BP/SF-1* (Mouse)	–224	->	+	+	+
6	*Ad4BP/SF-1* (Mouse)	–209	<-	+	-	+
7	*Ad4BP/SF-1* (Mouse)	-203	<-	+	-	no data
8	*Oxt* (Human)	-159	<-	-	+	+
9	*Cyp11B2* (Rat)	-324	->	-	+	+
10	*StAR* (Sheep)	–105	<-	+	-	+
11	*StAR* (Macaque)	-229	<-	+	-	+
12	*Ad* (Bovine)	-428	<-	-	+	+
13	*Cyp11B3* (Bovine)	-309	->	-	+	+
14	*Cyp11B1* (Sheep)	-337	->	-	+	+
15	*CYP17* (Porcine)	–51	->	+	+	+
16	*CYP17* (Porcine)	-140	<-	+	-	-
17	*HSD17BI* (Rat)	–84	<-	+	+	+
18	*StAR* (Bovine)	–104	<-	+	-	+
19	*StAR* (Bovine)	-244	<-	+	-	no data
20	*LH beta* (Porcine)	–114	<-	+	+	+
21	*Oxt* (Mouse)	-164	<-	-	+	+
22	*Oxt* (Rat)	-167	<-	-	+	+

* The position relative to the transcription start.

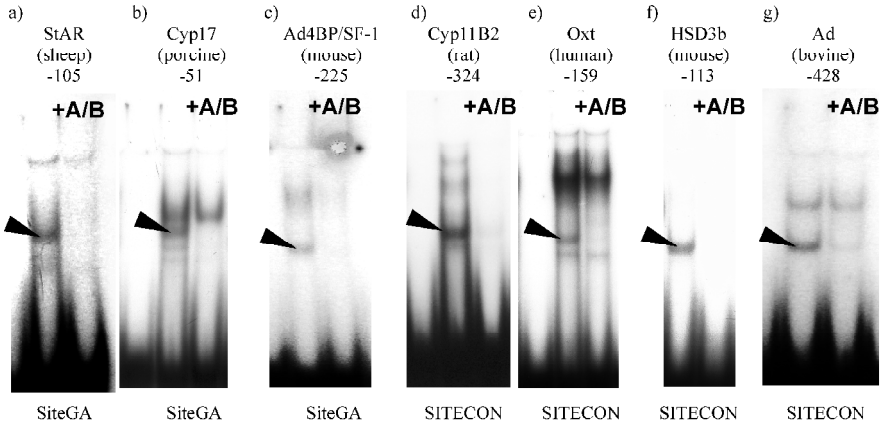


Fig. Experimental support for a number of SF-1 sites predicted by the SiteGA(a, b, c) and SITECON (d, e, f, and g) methods. Under the gene name is placed the site position relative to the transcription start. The shifted SF-1/DNA complex is indicated by arrow. Disappearance (or weakening) of the bands due to antibody (Upstate Biotechnology) against SF-1 (A/B) in the right lane confirms SF-1 binding to the site.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (No. 03-04-48506-a, 03-07-90181-в 03-04-48469-a), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Projects No. 119), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology". The authors are grateful to Pozdnyakov M.A., Proskura A.L., Lokhova I.V. for technical support and to Kolchanov N.A., Osadchuk A.V. for helpful discussions.

References

Busygina T.V., Ignatieva E.V., Osadchuk A.V. Consensus Sequence of Transcription Factor SF-1 Binding Site and Putative Binding Site in the 5;-Flanking Regions of Genes Encoding Mouse Steroidogenic Enzymes 3betaHSDI and Cyp17 // Biochemistry (Mosc). 2003. V. 68. P. 377–384.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acid Res. 2002. V. 30. P. 312–317.

Levitsky V.G., Ignatieva E. V., Busygina T.V., Merkulova T.I. Analysis of the context features of sf-1 binding site and development of a criterion for SF-1 regulated gene recognition by the SiteGA method // This issue. 2004.

Levitsky V.G., Katokhin A.V. Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis // In Silico Biol. 2003. V. 3. 0008. http://www.bioinfo.de/isb/2003/03/0008/

Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. SITECON—A tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition // Bioinformatics of genome regulation and structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London, 2004. P. 93–102.

Val P., Lefrancois-Martinez A.M., Veyssiere G., Martinez A. SF-1 a key player in the development and differentiation of steroidogenic tissues // Nuclear Receptor. 2003. V. 1. P. 8–45.

A NEW ALGORITHM FOR RECOGNIZING THE OPERON STRUCTURE OF PROKARYOTES

Ishchukov I.M.*, Likhoshvai V.A., Matushkin Yu.G.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: ishuk@bionet.nsc.ru.

Keywords: *prokaryotes, local complementarity, recognition algorithm, logical functions*

Summary

Motivation: So far, over 140 prokaryotic genomes have been sequenced. Bioinformatics tools are considerably more efficient for studying the structure of these genomes compared to classic genetic methods. One of the goals is to recognize the operon structure of the genome by computer methods.

Results: A new original algorithm for recognizing the operon structure was developed; this algorithm is based on analysis of local self-complementarity of sequences of genes and logical recognition functions. The recognition accuracy of cistron position in the operon of *E. coli* and related organisms amounts to ~70 %; of *B. subtilis,* ~65 %. The algorithm operation speed is considerably higher compared with the analogous methods.

Availability: http://www.bionet.nsc.ru/bgrs2004/.

Introduction

Availability of the complete genomic sequences underlies a global approach to studying the structure of these genomes by bioinformatics tools. Research into the operon structure of prokaryotic genomes is a necessary step in this direction.

The state-of-the-art knowledge on regulation and functions of *E. coli* K12 genes is incomparable with the analogous data concerning any other organism. *E. coli* is a perfect model organism for testing the tools of computer genomics. Upon studies involving *E. coli*, the corresponding methods could be used for a more detailed research of other organisms.

The majority of methods for determining the operon structure developed so far utilize the information on specific sequence motifs. We developed a new algorithm that makes use only of the primary structure of the genome and standard information on protein-coding sequences. The algorithms may be applied to prediction of the operon structure of insufficiently studied organisms, assisting in turn in prediction of regulatory interactions at the level of transcription initiation.

Materials, methods, and algorithms

Various functions of nucleotide sequence compositions can be constructed basing on mRNA or DNA sequences. The local complementarity index (LCI) is a measure of the number and energy of the local secondary structures (hairpins; Likhoshvai, Matushkin, 2002). This measure was successfully used when studying the factors influencing the efficiency of gene expression.

In this work, we calculated LCI for each nucleotide; thus, this index reflects the energy of the totality of all the possible secondary structures involving a particular nucleotide. As a result, we construct the LCI profile of a sequence and use the pattern of this profile as a characteristic feature.

All the genes were extracted from the complete genomic sequence of *E. coli,* available in GenBank. The genome was divided into operons using the algorithm based on accounting the intergene distance (Salgado *et al*., 2000). As a limit, we selected 40 nucleotides (the optimal distance for *E. coli*). When studying the profiles, the sequences were aligned according to start and stop codons. Overall, 17 groups of genes were formed depending on their location in the operon,

and LCI profiles were constructed for each group. Each gene group displays individual profile; however, certain regions of the profiles from different groups are similar. Shown in Fig. 1 is the averaged LCI profile constructed for all the *E. coli* genes.

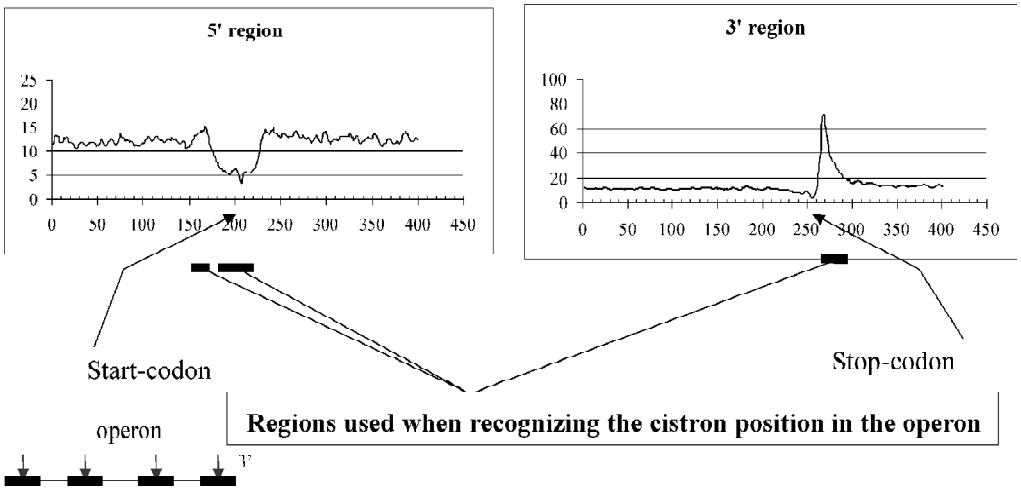


Fig. 1. The averaged LCI profile constructed *for all the E. coli* genes (the sample of 4284 genes): the ordinate, the LCI value calculated for the corresponding nucleotides.

As is evident from Fig. 1, the LCI profile exhibits specific features in the vicinity of the start and stop codons, and this can be used for recognition of a cistron in the operon. The following fragments were used in the algorithm: relative to the start codon, (–15, +25), (–40, –30), and (+50, +175)—(–175, –50; taking into account the difference in LCI levels; relative to the stop codon (position 0), (0, +50) and (+75, +175)—(–150, –50; taking into account the difference in LCI levels at last two sites).

In particular, the LCI profile calculated for the cistrons that are not the first or the last (Fig. 2) differs considerably from both the averaged profile and the profiles of the other groups of cistrons. These differences allow for predicting to which group a cistron belongs basing on its LCI profile. Logical decision functions are among most suitable solution classes for such prediction problems (Lbov, Startseva, 1999; Lbov, 2000). The decision function is constructed basing on a data table (training sample). The problem of finding the best decision function is a complex extreme problem. One of the procedures for searching for the best decision function is the LDR (logical decision rule) algorithm (Lbov, Startseva, 1999; Lbov, 2000). In our case the learning sample was a randomly chosen subset of known cistrons.

The LDR algorithm performs a stepwise division of classes of objects into two subclasses that differ maximally from one another. As a result, we obtain a set of rules allowing an unknown object to be attributed to a particular class. The rules are specified in a certain order. Branching of the tree is performed in a stepwise manner.

Upon constructing, the tree is used to predict objects belonging to an unknown class. Each object falls to the end node upon successive checking that the conditions are met and a certain value is ascribed to the object. In our case cistrons are the objects.

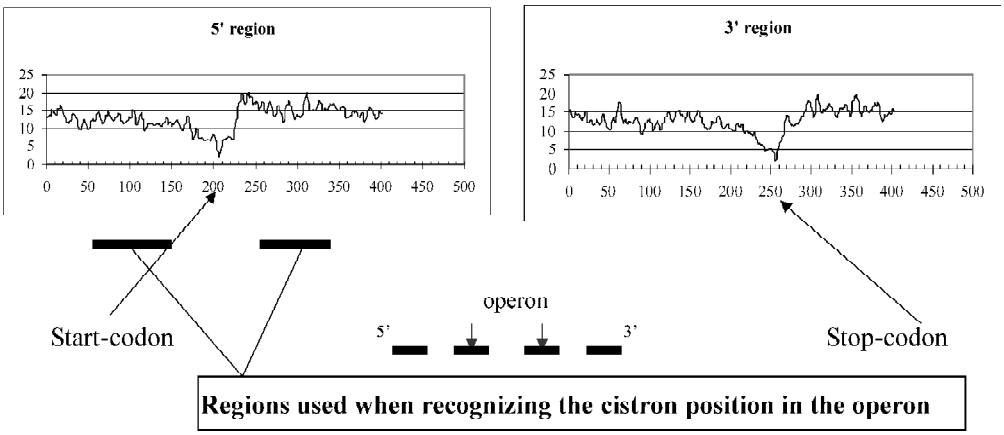


Fig. 2. The averaged LCI profile constructed for the cistrons of *E. coli* operons *that are not the first or the last* (the sample contains 776 genes).

Results and Discussion

The LDR algorithm was realized in C++. Sums of LCI values at the fragments indicated above were used as the variables characterizing the object—cistron, i.e., all the variables are real. The training tables were created using the samples of experimentally confirmed operons available at the site www.ecocyc.org (692 operons). Four samples were formed, namely: 1) first cistrons of operons; 2) cistrons of operons that are not the first; 3) last cistrons of operons; and 4) cistrons of operons that are not the last.

The monocistronic operons were included into both sample 1 and sample 3.

Two recognition trees were constructed: one for recognizing the first cistron of the operon; the second for recognizing the last cistron. Upon running the algorithm, each cistron falls into one of the following classes: 1) the first cistron of a multicistronic operon; 2) the last cistron of a multicistronic operon; 3) an inner cistron of an operon; and 4) a monocistronic operon.

These two trees have different logical structures (their nodes contain different predicates and the branching patterns are different) and "operate" with different accuracies. The gene regions putatively crucial for determining the locations of genes in the operon were chosen for constructing these trees. Then, they were optimized and the key regions were finally detected. The total recognition accuracy for each tree was calculated as a sum of the probabilities of the cistrons to be correctly determined by the end nodes. The first cistrons are recognized with an accuracy of 67 %; last cistrons, of 69 %.

The recognition algorithm was applied to determining the operon structure of *Bacillus subtilis,* a bacteria related to *E. coli*. Its operon structure is also sufficiently well studied, also in less detail compared with *E. coli*. The general pattern of LCI profiles constructed for *B. subtilis* were similar to those of *E. coli*.

The boundaries for calculation of the profile—a drop in the vicinity of the start codon and a peak near the stop codon—remained unchanged. Initially, the trees obtained using *E. coli* were applied to *B. subtilis* without adaptation. The results obtained amounted to 59 and 66 % for the 5'- and 3'-regions, respectively. When the recognition algorithm was adapted to *B. subtilis,* the peak in profile in the vicinity of the start codon appeared shifted. When the separation of *B. subtilis* at the 5'-end of the intervals used for constructing the profile, the effectiveness amounted to 62 %.

The results obtained for *B. subtilis* using this algorithm suggest the inference that the behavior of

the LCI profiles constructed for related organisms is similar, allowing the algorithm to be applied without considerable adaptations.

The algorithm developed can be used for an approximate prediction of the position of any cistron in the operon. Its effectiveness is similar to the algorithms described in other works (Salgado *et al*., 2000, Yada *et al*., 1999; etc.). However, the algorithm developed requires several dozens minutes for its operation using a conventional PC, whereas the algorithm developed by Yada *et al*. (1999) runs several hours at a supercomputer. Several reasons may cause errors in determining the location of a cistron. For example, the direction of a considered cistron relative to the neighboring cistrons is not taken into account completely. However, this factor might be most important for determining the cistron position.

Development of new, more general criteria for construction of recognition trees will form the background for development of an algorithm recognizing the operon structure of insufficiently studied organisms.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 02-04-488802 and 03-04-48829); Russian Ministry of Industry, Sciences, and Technologies (grants Nos. 43.106.11.0011; 43.073.1.1.1501); SB RAS Integration project No. 148, RAS Presidium Program "Molecular and Cellular Biology" (project No. 10.4) and "Origin and Evolution of biosphere" (state contract № 10002-251/П-25/155-270/200404-082).

References

Lbov G.S. The Theory and Methods for Construction of Functions for Sample Recognition. Novosibirsk: Novosibirsk State University, 2000.

Lbov G.S., Startseva N.G. Logical Decision Functions and Problems of Statistical Stability of Solutions. Novosibirsk: Institute of Mathematics, 2000.

Likhoshvai V.A., Matushkin Yu.G. Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy // FEBS Letters. 2002. V. 516(1/3). P. 87–92.

Salgado H., Moreno-Hagelsieb G., Smith T., Collado-Vides J. Operons in *Escherichia coli*: Genomic analyses and predictions // PNAS. 2000. V. 97, N 12. P. 6652–6657.

Yada T., Nakao M., Totoki Y., Nakai K. Modelling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models // Bioinformatics. 1999. V. 15, N 12. P. 987–993.

ANALYSIS OF OLIGONUCLEOTIDE COMPOSITION IN DNA OF *E. COLI* GENOME AND PROMOTER SITES

Kamzolova S.G.*, Sorokin A.A., Dzhelyadin T.R., Osypov A.A., Beskaravainy P.M.

Institute of Cell Biophysics of RAS, Pushchino Moscow region, Russia
* Corresponding author: e-mail: ao@icb.psn.ru; kamzolova@icb.psn.ru

Keywords: *E. coli DNA, nucleotide sequence, promoter sites, oligonucleotide composition, RNA polymerase, recognition*

Summary

Motivation: Recent identification of some noncanonical determinants in *E. coli* promoters has opened a field for a search of new candidates for this role. As suggested and shown here, a large-scale comparative analysis of oligonucleotide composition of the complete sequence of *E. coli* genome and its promoter sites has much potential for yielding information about all kinds of nucleotide blocks preferred by promoters that might contribute to promoter recognition.

Results: The content of all kinds of hexa-, hepta- and octonucleotides in the complete *E. coli* genome and in 359 promoters was determined. A nonrandom distribution of oligonucleotides was found both in the complete genome and in the promoters. The complete genome and the promoters were shown to differ essentially in their oligonucleotide sets. 542 different hexanucleotides are twofold more frequent in the promoters than in the chromosome. Hexanucleotides most preferred by promoters are discussed as possible candidates for the role of new promoter determinants.

Availability: DNA hexanucleotide analysis software is available at request to academic users (lptolik@icb.psn.ru)

Introduction

There are about 4000 promoters in genome of *E. coli* and its related bacteriophages that are recognized by RNA polymerase ($E\sigma^{70}$). The promoters have been shown to be considerably varied in their nucleotide sequences. Statistical analysis of nucleotide sequences for all known promoters has revealed two homologous hexamer motifs, 5'TATAAT3' and 5'TTGACA3' centered around position -10 and -35, respectively (Harley, Reynolds, 1987). The functional role of these hexanucleotides as universal promoter determinants involved in recognition of RNA polymerase has been proved. However, even in these conserved regions nucleotide sequences differ essentially in individual promoters containing, as a rule, only 7.9 canonical nucleotide pairs from 12. Therefore, the precise identification of promoters in nucleotide sequence of *E. coli* genome on the basis of the canonical consensus sequences appeared to be impossible.

Taking into account the sequence diversity of the promoters, a wide range of their functional activities and their differential response to the same physiological signals, it was suggested that some noncanonical specific determinants should be involved in the process of RNA polymerase interaction with different promoters and their groups. Some particular promoter groups have been recently shown to be described by specific noncanonical consensus sequences, possible candidates for the role of new specific promoter determinants (reviewed in Kamzolova, 1995). To continue the search for new noncanonical sequence elements in promoters, here, a large-scale comparative analysis of oligonucleotide composition was undertaken for the complete sequence of *E. coli* genome and its promoter sites.

Methods and Algorithms

The complete sequence of *Escherichia coli* K-12 genome containing 4639221 bp was taken from GenBank U00096. A set of 359 promoters designated on *E. coli* genome map as experimentally identified was chosen for this study.

All nucleotide sequences were compiled and analyzed with the computer programs of Sorokin (Sorokin, 2001).

Results

E. coli genome contains 4288 genes combined into 2580 operons that are controlled by ~ 4000 promoters. 359 promoters interacting with RNA polymerase -Eσ^{70} were identified and localized on the genome map.

To determine a set of preferred nucleotide blocks selected by *E. coli* RNA polymerase in evolution process, the presence of the full set of hexa-, hepta- and octonucleotides was tested in sequences of 359 promoters and *E. coli* complete genome. The results obtained are presented in Table 1 and Table 2.

Table 1. Distribution of oligonucleotides in E. coli chromosome and its promoters

n-Oligonucleotides		6-membered	7-membered	8-membered
The total number of n-oligonucleotides		4096	16384	65536
The number of n-oligonucleotides forbidden in chromosome		0(0%)	1(0.006%)	173(0.26%)
The number of n-oligonucleotides forbidden in promoters		111(2.71%)	5131(31.3%)	46796(71.4%)
The number of n-oligonucleotides	z>2	542(13.2%)	3237(19.8%)	13332(20.3%)
	z>5	35(0.85%)	470(2.9%)	4990(7.6%)
	z>10	5(0.12%)	89(0.54%)	1481(2.3%)

$z=\omega_p/\omega_c$, where ω_p and ω_c are the frequencies of oligonucleotide occurrence in promoters and chromosome, correspondingly.

Discussion

The data obtained indicate that *E. coli* DNA is characterized by a nonrandom distribution of oligonucleotides (Table 1 and 2). There are many preferred oligonucleotides (Table 2, columns ω_δ and ω_c) as well as some forbidden blocks (Table 1) in it. This applies both to the complete nucleotide sequence of the genome and to its promoter sites. It is notable that a set of oligonucleotide preferred in the promoters is distinctly different from that in the chromosome. Table 2 shows 15 major hexanucleotides that are predominant in the chromosome (ω_c-column) or in the promoters (ω_p-column). The set of the predominant hexanucleotides for the chromosome is marked by GC-enriched blocks, whereas AT-enriched sequences predominate in the promoters. Short oligoA-oligoT runs similar to those found in the promoters are known to induce DNA curvature, a structural feature that might contribute to promoter recognition (Shpigelman *et al*., 1993). In addition, AT-enriched sequences are found to be "easily melting" (thermodynamically less stable or non-stable) sites that are also predisposed to interaction with RNA polymerase (Kamzolova, Postnikova, 1981). Thus, it appears reasonable to suggest that the possible functional role of some oligonucleotides as promoter determinants can be expressed through their physical properties (such as curving and bendability, thermostability, or electrostatic properties). 542 different hexanucleotides are twofold more frequent in the promoters than in the complete genome

(Table 1, row 4 - z>2). These hexanucleotides are very large in number to consider all of them as sequence-specific promoter determinants. It is natural to suppose that some of these hexanucleotides have no functional meaning, some of them can be involved in the interaction with RNA polymerase due to their specific physical properties and only a few of them may contain functional information hidden away in the base sequence of these blocks. Hexanucleotides listed in the third column of Table 2 as being most preferred in promoter sites seem to be very promising for experimental studies as candidates for the role of new promoter determinants.

Table 2. Characterization of *E. coli* chromosome and its promoter sites by hexanucleotides composition

	ω_c	ω_p	$z=\omega_p/\omega_c$
1	CGCCAG	TTTTTT	TACTAG
2	CTGGCG	ATTTTT	ACTAGA
3	GCCAGC	AAAAAA	ACTAGT
4	GCTGGC	TTTTTA	AACTAG
5	CCAGCG	TTTTTC	*TAGAAT*
6	CGCTGG	TTTTCA	ACTTGT
7	CCAGCA	TTTTAT	CTAGAG
8	CAGCGC	TTTATT	CTAGAC
9	TGCTGG	AAAAAT	CTAGTG
10	GCGGCG	TAAAAT	**TATAAT**
11	CGCCGC	ATAAAA	*TATACT*
12	CAGCAG	AAAATT	ACACTT
13	TCGCCA	AAATTT	CTAGGT
14	GCGCTG	TTTTGT	TAGACT
15	CTGGCA	CATTTT	TACACT

Hexanucleotides are arranged in decreasing order according to frequencies of their occurrence in chromosome (ω_c), in promoters (ω_p) and promoters to chromosome occurrence frequencies ratio ($z=\omega_p/\omega_c$).

Acknowledgements

This work was supported by Russian Foundation for Basic Research (grants: RFBR-naukograd 04-04-97275 and RFBR-04-04-49635).

References

Harley C.B., Reynolds R. Analysis of Escherichia coli promoter sequences // Nucl. Acids Res. 1987. V. 15. P. 2343–2361.

Kamzolova S.G. Classification approach in studying RNA polymerase-promoter code // Biochimiya. 1995. V. 61. P. 1128–1131.

Kamzolova S.G., Postnikova G.B. Spin-labeled nucleic acids // Quart. Rev. Biophys. 1981. V. 14. P. 223–288.

Shpigelman E.S., Trifonov E.N., Bolshoy A. Curvature: software for the analysis of curved DNA // Comput. Appl. Biosci. 1993. V. 9. P. 435–440.

Sorokin A.A. Functional analysis of *E. coli* promoter sequences. New promoter determinants. Ph. D Thesis, Pushchino. 2001.

ELECTROSTATIC PROPERTIES OF *E.COLI* GENOME DNA

Kamzolova S.G., Sorokin A.A., Dzhelyadin T.R., Beskaravainy P.M.*, Osypov A.A.

Institute of Cell Biophysics of RAS, Pushchino, Moscow region, Russia
* Corresponding author: e-mail: bpm@icb.psn.ru; kamzolova@icb.psn.ru

Keywords: *E. coli genome, promoters, coding region, electrostatic potential distribution*

Summary

Motivation: Distribution of electrostatic potential around nucleotide sequences is one of fundamental characteristics of DNA contributing to its recognition by DNA-binding proteins. Analysis of electrostatic properties of natural DNAs had to await the development of appropriate calculation methods for long nucleotide sequences. A method recently proposed in our work (Kamzolova *et al*., 2000) satisfies the requirement thus opening a promising means for studying electrostatic properties complete genomes and their different regions. Here, the method was used for analysis of electrostatic potential of *E. coli* genome.

Results: Distribution of electrostatic potential of the complete sequence of *E. coli* genome was calculated. It is found that DNA is not a uniformly charged molecule. There are some local inhomogeneities in its electrostatic profile which correlate with promoter sequences. These characteristic variations of electrostatic potential of DNA may be involved in RNA-polymerase-DNA recognition.

Availability: Electrostatic potential distribution analysis software is available at request to academic users (lptolik@icb.psn.ru).

Introduction

One of fundamental physico-chemical characteristics of any macromolecule, which affects its interaction with different ligands, is the distribution of electrostatic potential around it. DNA is a highly charged polyelectrolite and therefore its electrostatic potential may be one of the main features recognized by DNA-binding proteins. Really, electrostatic interactions between promoter DNA and *E. coli* RNA-polymerase ($E\sigma^{70}$) has been recently shown to be of considerable importance in regulating promoter function (Kamzolova *et al*., 2000). Electrostatic characteristics of promoter DNA have been suggested to be a new promoter determinant marked by its relative independence from promoter nucleotide sequence. The important role of electrostatic interactions in the multi-step process of protein-DNA recognition has also been shown for some other DNA-binding proteins (Misra *et al*., 1994; Fogolary *et al*., 1997; Labeots, Weiss, 1997) It is not surprising, then, that theoretical analysis of electrostatic potential distribution around DNA molecule and its fragments containing protein-binding sites is one of the pressing areas in modern research of protein-DNA recognition coding. Recently we have proposed a simplified method for calculation of electrostatic potential distribution for long nucleotide sequences (Kamzolova *et al*., 2000; Sorokin, 2001). Here the method was used for theoretical analysis of electrostatic properties of the complete sequence of E. coli genome. As an example to illustrate a functional meaning of the information hidden away in the electrostatic map of the genome, a comparative analysis of electrostatic patterns of promoter and nonpromoter DNA sites was made.

Methods

The complete sequence of *E. coli* K-12 genome was taken from GenBank (accession number U00096). The electrostatic potential around double-helical DNA molecule was calculated by the Coulombic method (Kamzolova *et al*., 2000) using the computer program of Sorokin A. (Sorokin, 2001).

Results and Discussion

When considering the prospects in studying electrostatic properties of a complete molecule of natural DNAs it should be emphasized that until recent years such analysis has been hampered due to considerable difficulties of theoretical calculations of electrostatic potentials for long DNA fragments. An accurate calculation can be carried out only for short DNA sequence not more then 30–40 nucleotide pairs using nonlinear Poisson – Boltzman equation (Jayaram *et al*., 1989). Recently we have proposed a simplified method for calculation of electrostatic potential distribution based on Coulomb's law which can be used for long fragments of double – helical DNA including complete genomes (Kamzolova *et al*., 2000; Sorokin, 2001). Though the method cannot be applied for rigorous treatment, it is quite suitable for qualitative analysis of electrostatic map of a natural DNA molecule and for comparative studies of its different parts.

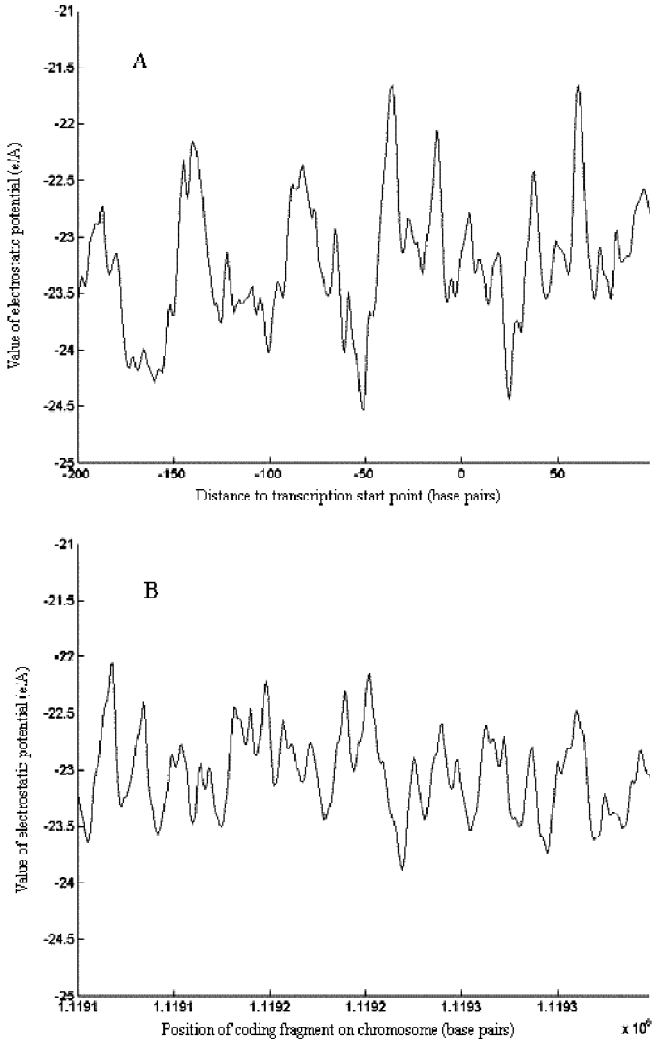


Fig. Electrostatic potential distribution for promoter (A) and coding region (B) in *E. coli* genome.

Using this method, calculation of electrostatic potential distribution was performed for complete nucleotide sequence of *E. coli* genome containing 4639221 base pair. The results obtained representing the profile of electrostatic potential distribution around the complete genome are in our database (kamzolova@icb.psn.ru)

The possibility of extracting some functional information from the electrostatic map of *E. coli* genome can be demonstrated by the example provided by a large-scale analysis of electrostatic patterns of promoter and nonpromoter DNA sites. Electrostatic profiles of 359 promoters identified in *E. coli* genome as well as of their nearby coding sequences were analyzed by the presence of peaks and valleys as well as by their arrangement and values. Figure shows some representative examples of electrostatic patterns for promoters (A) or DNA coding regions (B). It is found that coding regions are characterized by more homogeneous distribution of electrostatic potential, whereas local inhomogeneities with the most positive and negative areas correspond to promoter sites. It should be noted that individual promoters discussed here vary in the design of their electrostatic profiles but all of them, in contrast to coding regions, are characterized by inhomogeneous complex-shaped patterns. These characteristic variations of electrostatic potential of DNA may be related to RNA polymerase – DNA recognition by specifying promoter sites as electrostatic traps or barriers for the enzyme. In addition, alternating areas of negative and positive potential in promoter sites may enforce charged RNA polymerase molecules to orient properly relative to the transcription start point.

Thus, DNA electrostatic component may be one of the determining factors allowing RNA polymerase to identify promoter sites in genomes.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grants: RFBR naukograd-04-04-97275 and RFBR 04-04-49635).

References

Fogolari F., Elcock A.H., Esposito G., Viglino P., Briggs J.M., McCammon J.A. Electrostatic effects in homeodomain-DNA interactions // J. Mol. Biol. 1997. V. 267(2). P. 368–381.

Kamzolova S.G., Sivozhelezov V.S., Sorokin A.A., Dzhelyadin T.R., Ivanova N.N., Polozov R.V. RNA polymerase—promoter recognition. Specific features of electrostatic potential of “early” T4 phage DNA promoters // J. Biomol. Struct. Dyn. 2000. V. 18(3). P. 325–334.

Labeots L.A., Weiss M.A. Electrostatics and hydration at the homeodomain-DNA interface: chemical probes of an interfacial water cavity // J. Mol. Biol. 1997. V. 269(1). P. 113–128.

Misra V.K., Hecht J.L., Sharp K.A., Friedman R.A., Honig B. Salt effects on protein-DNA interactions. The lambda cI repressor and EcoRI endonuclease // J. Mol. Biol. 1994. V. 238(2). P. 264–280.

Sorokin A.A. Functional analysis of E. coli promoter sequences. New promoter determinants. Ph.D. Thesis. Pushchino, Institute of Theoretical and Experimental Biophysics RAS. 2001.

MOLECULAR PALEONTOLOGY OF DNA TRANSPOSONS IN EUKARYOTIC GENOMES

Kapitonov V.V.*, Jurka J.

Genetic Information Research Institute, Mountain View, California, USA
* Corresponding author: e-mail: vladimir@ulam.girinst.org

Keywords: *eukaryotic genome, superfamilies of DNA transposons, repetitive elements, molecular paleontology, computational biology*

Summary

Motivation: In eukaryotes, interspersed repetitive elements constitute 3–60 % of the genome. Usually, these elements are relics of transposable elements. Given the growing list of examples showing a significant impact of transposable elements on evolution and function of the genome, classification of repetitive elements based on their biological properties is an important task. Moreover, novel classes of transposable elements and new biological processes can be discovered based on systematic computational studies of repetitive elements.

Results: Here we show that the entire variety of all DNA transposons in eukaryotes can be described as a set of elements that belong to only a few superfamilies. Based on computational studies reported here novel superfamilies of DNA "cut and paste" transposons (*Harbinger*, *Transib*, *Mirage*) and "rolling-circle" DNA transposons (*Helitrons*) were discovered and characterized. We describe major properties of these superfamilies, including proteins necessary for transpositions, target-site specificity, and structure of transposon termini.

Availability: Sequences of all transposable elements described here are available at http://www.girinst.org/Repbase_Update.html

Introduction

Transposable elements (TEs) constitute a large fraction of the genomes of eukaryotes. Based on their transposition mechanism, they have been divided into two classes. Transposition of Class 1 elements (retrotransposons) is catalyzed by reverse transcriptase and endonuclease/integrase encoded by some of them. Class 1 elements RNA expressed in the host cell is reverse transcribed, and subsequently the resulting cDNA copies can be integrated into the host genome. The reverse transcription and integration are catalyzed by the element-encoded reverse transcriptase and endonuclease/integrase, respectively. Transposition of Class 2 elements (DNA transposons) is catalyzed by element-encoded transposases (TPases). It was thought that Class 2 elements proliferate via so-called "cut and paste" transpositions: they are excised (cut) from the original genomic location and inserted (pasted) into a new site. These reactions are catalyzed by the transposase that binds termini of a transposon and introduces DNA nicks. DNA transposons identified previously in eukaryotic genomes have been assigned to six superfamilies: *Tc1/mariner*, *hAT*, *MuDR*, *En/Spm*, *piggyBac*, and *P* (Fig. 1). Each superfamily includes numerous diverse families composed of autonomous and nonautonomous elements, whose transposition is catalyzed by superfamily specific transposases. While an autonomous element encodes a complete set of enzymes characteristic of its family and is self-sufficient in terms of transposition, a nonautonomous element transposes by borrowing the protein machinery encoded by its autonomous relatives.

Methods and Algorithms

Computational analysis permits efficient reconstruction of ancient TEs from their incomplete or mutated copies scattered in the genome that we consider as genomic DNA fossils. Main methods

used in our studies were described previously (Kapitonov, Jurka, 1999, 2000, 2003, 2004). DNA and protein similarities were identified by using stand-alone NCBI BLAST, WU BLAST, PSI-BLAST, and CENSOR. Exons and introns were predicted by Genescan and Fgenesh. Multiple alignments were produced by CLUSTAL-W, T_Coffee, and tools designed in GIRI (implementations of the Smith-Waterman algorithm and SWAT program).

Results and Discussion

Harbinger is the first superfamily of DNA transposons discovered based on computational studies (Kapitonov, Jurka, 1999, 2004). Relics of *Harbinger* transposons are present in the genomes of plants, mammals, vertebrates, fungi, insects, and diatoms. Recently, transpositions of nonautonomous *Harbingers* have been detected in the rice genome. The autonomous *Harbingers* encode two proteins: a ~400-aa *Harbinger* TPase and ~200-aa DNA-binding protein that includes the conserved SANT/myb/trihelix motif. The *Harbinger* TPase is distantly related to TPases encoded by the IS5-like group of bacterial transposons, such as IS5, IS112, and ISL2. Usually, integration of a *Harbinger* transposon is accompanied by a 3-bp duplication of the TAA or TTA target site. Some *Harbingers* found in the zebrafish genome are characterized by a striking preference for a 17-bp target site never seen previously in any other DNA transposons (the AAAACACCWGGTCTTTT consensus sequence). Although mammalian genomes do not harbor recognizable relics of *Harbinger*s, we identified a widely expressed HARBI1 gene encoding a 350-aa protein entirely derived from a *Harbinger* TPase some 450 million years ago. The HARBI1 protein is highly conserved in vertebrates, including human, rat, mouse, cow, pig, chicken, frog, and bony fish. For example the human HARBI1 protein diverges only 5 % and 40 % from its counterparts in rat and fish, respectively. All conserved motifs detected in the *Harbinger* TPases are also well preserved in the HARBI1 proteins. Therefore, the HARBI1 proteins are expected to serve as nucleases important for bony vertebrates. Most likely, the HARBI1 proteins are involved in genome rearrangements.

We also discovered a novel superfamily of DNA transposons called *Transib* (Kapitonov, Jurka, 2003) that populate the fruit fly and mosquito genomes. These transposons use a ~700-aa *Transib* TPase that is not similar to any known proteins, and are characterized by 5-bp target site duplications (TSDs). Based on the conservation profile of the *Transib* TPase, we identified a putative D(34)D(35)E triad, which serves putatively as a catalytic core involved in DNA cleavage during transposition.

Mirage is the third novel superfamily of eukaryotic "cut and paste" DNA transposons discovered based on computational studies. So far, *Mirage* transposons have been identified in the nematode genome only. Autonomous *Mirage* transposons encode the *Mirage* TPase, which is not similar to TPases encoded by TEs that belong to known superfamilies. We classified this protein as a novel TPase due to its presence in a few highly diverged families of *Mirage* TEs. The *Mirage* transposons are flanked by conserved terminal inverted repeats and they generate 2-bp TSDs upon their integration into the host genome.

As a result of systematic computational studies we discovered a novel category of eukaryotic DNA transposons, called *Helitrons*, which transpose via replicative rolling-circle transposition (Kapitonov, Jurka, 2001). *Helitrons* are present in the genomes of plants, insects, fish, fungi, and worms. In some species, like *A. thaliana* and *C. elegans*, they make up 2 % of the genome. Autonomous *Helitrons* encode the ~1500-aa Rep/Hel protein composed of the Rep and Hel conserved domains. The Rep domain spans a ~160-aa region composed of the "two-His" (E-FYW-Q-K-R-G-LAV-PV-H-X-H) and "KYK" ("Yg-LVW-FAT-Kq-Y-X-X-K) motifs separated by ~130 aa. These two motifs are known to be conserved in Rep-like prokaryotic proteins encoded by various plasmids and single-stranded DNA viruses that employ rolling-circle DNA replication. Most importantly, these Rep proteins perform both cleavage and ligation of DNA during rolling-circle replication. The ~500-aa Hel domain is a helicase that belongs to the SF1 superfamily of DNA helicases. *Helitron* is the only class of transposons in eukaryotes that integrate into the genome without introducing TSDs. Usually,

the *Helitron* integration occurs precisely between host A and T nucleotides. Surprisingly, *Helitrons* do not have terminal inverted repeats, which are typically present in other DNA transposons. Instead, *Helitrons* have conserved 5'-TC and CTRR-3' termini. They also contain a ~18-bp hairpin separated by 10–12 nucleotides from the 3' end. Presumably, the hairpin serves as the terminator of rolling-circle replication, which is believed to be a mechanism of *Helitron*'s transposition. So far, only *Helitrons* found in the *Aspergillus nidulans* genome do not contain the 3' hairpin.

Although no active *Helitrons* have been isolated so far and studied experimentally, the main features of *Helitron* transposition can be predicted *a priori* based on the structural invariants detected in different *Helitrons* and known properties of bacterial rolling-circle replicons. The current model for *Helitron* transposition/replication involves a few basic stages: (a), it starts from a site-specific Rep-encoded nicking of the transposon plus strand; (b), A free 3'-OH end of the nicked plus strand serves as a primer for leading-strand DNA synthesis facilitated by the *Helitron* helicase and some host replication proteins, including DNA polymerase and RPA-like single-stranded DNA-binding proteins. The newly synthesized leading plus strand remains covalently linked to the 3'-OH end of the parent plus strand during the continuous displacement of its 5'-OH end. When the leading strand makes a complete turn, Rep catalyzes a strand-transfer reaction followed by release of a single-stranded DNA intermediate, the parent minus strand, and a double-stranded DNA *Helitron* composed of both the parental plus and a newly synthesized strand.

Most *Helitrons* are nonautonomous elements. They share common termini and other structural hallmarks with autonomous *Helitrons*. Some families of nonautonomous *Helitrons* in plants, nematode, and mosquito include identical elements, and the genomes do not harbor autonomous elements that are direct relatives of the nonautonomous elements. Therefore, we do expect that corresponding autonomous elements are active transposons that are not fixed in the genome yet.

Another interesting feature of *Helitrons* is their ability to transduce host genes. For example plant *Helitrons* encode replication protein A (RPA)-like proteins, clearly derived from RPA encoded originally by the host genome. Given the conservation of RPA in *Helitrons*, this protein should be involved in transposition of *Helitrons*, presumably as a single-stranded DNA-binding protein. Some families of *Helitrons* present in the zebrafish genome carry endonuclease derived from CR1-like non-LTR retrotransposons. Some *Helitrons* detected in the corn genome harbor exon-intron-coding portions of several host genes. Therefore, *Helitrons* can function as a powerful tool of evolution. They are capable of recruiting host genes, of modifying them to an extent that is unattainable by standard Mendelian processes, and of multiplying them in the host genome.

Table. Superfamilies of eukaryotic DNA transposons

Superfamily	Host species	Related bacterial TPases	TSDs (bp)
hAT	very wide spectrum	–	8
Tc1/mariner	very wide spectrum	IS630	2, usually TA
En/Spm	plants, fish	–	3
MuDr	diatom, plants, worms, mammals	IS256, IS1016	8–10
piggyBac	insects, mammals, vertebrates	–	4, usually TTAA
P	insects, vertebrates, mammals	–	7–8
Mirage	worms	–	2
Harbinger	very wide spectrum	IS5, IS112	3, usually TWA
Transib	insects	–	5
Helitron	plants, worms, insects, vertebrates, fungi	IS91	–

Acknowledgements

This work was supported by the National Institutes of health grant 2 P41 LM06252-04A1.

References

Kapitonov V.V., Jurka J. Molecular paleontology of transposable elements from *Arabidopsis thaliana* // Genetica. 1999. V. 107. P. 27–37.

Kapitonov V.V., Jurka J. Rolling-circle transposons in eukaryotes // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 8714–8719.

Kapitonov V.V., Jurka J. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome // Proc. Natl Acad. Sci. USA. 2003. V. 100. P. 6569–6574.

Kapitonov V.V., Jurka J. *Harbinger* transposons and an ancient HARBI1 gene derived from a transposase // DNA Cell Biol. 2004 (in press).

ANALYSIS OF NUCLEOSOME FORMATION POTENTIAL AND CONFORMATIONAL PROPERTIES OF HUMAN J1-J2 TYPE ALPHA SATELLITE DNA

Katokhin A.V.*, Levitsky V.G., Oshchepkov D.Yu., Poplavsky A.S., Furman D.P.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: katokhin@bionet.nsc.ru

Keywords: *alpha satellite DNA, nucleosome positioning, DNA conformational properties*

Summary

Motivation: The structure-forming function of alpha satellites in arrangement of the centromeric and pericentromeric heterochromatin implies the presence of certain contextual and conformational signals (codes) for compacting DNA of these regions into nucleosomes and chromatin structures of higher level. However, this aspect of informational content of primary DNA sequences of alpha satellites yet requires further studies.

Results: Computer analysis of nucleosome formation potential (NFP) was performed using a sample of J1–J2 type alpha satellites from the human genome. Several regions with the context favorable for several variants of nucleosome positioning were detected. Statistical analysis of distribution of DNA conformational properties, in particular, the property Wedge, demonstrates a superposition of the context-dependent and CENP-B-dependent nucleosome positionings.

Availability: The corresponding software packages are available at http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/ (the method RECON) and http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/ (the method SITECON).

Introduction

Tandemly repeated dimers of alpha satellites of the J1–J2 types whereon chain of nucleosomes are positioned represent the centromere-forming elements of nine human chromosomes. Each monomer has a length of ~171 bp and permits positioning of only one nucleosome (146 bp) with a short linker region, resulting in a supercompact DNA packaging in centromeric and pericentromeric regions (Gilbert and Allan, 2001). Characteristic of the monomers of J2 type is the presence of B box—the site for specific binding the centromeric protein CENP-B (Yoda *et al*., 1998). The event of CENP-B binding to B box determines an unambiguous nucleosome positioning within the dimer, as was demonstrated in *in vivo* (Ando *et al*., 2002) and *in vitro* experiments (Yoda *et al*., 1998). The experiments on *in vitro* reconstitution of nucleosomes in the absence of CENP-B protein showed that several equivalent translational positions for the entire chain of nucleosomes with retained fixed internucleosome interval were realized (Yoda *et al*., 1998). This fact suggests a sequence-dependent nucleosome positioning; however, the nature of this dependence is yet vague.

In this work, computer analysis of distribution of contextual and conformational nucleosome positioning signals in J1–J2 type dimers was performed. The results obtained agree well with the data on experimental nucleosome mapping.

Methods and Algorithms

Samples of actual J1 and J2 dimers differing in the context from the consensus not more than by 5% were used in the work. The sequences were extracted from the following genomic contigs: BX284928, AC069355, AC135046, AB005791, NC_000007, BX322613, AADD01123003, M58446, and AC135053.

The following sequence characteristics essential for nucleosome positioning were assessed: NFP by the method RECON (Levitsky *et al*., 2001) and distribution of 38 statistically significant conservative context-dependent DNA conformational and physicochemical properties by the method SITECON (Oshchepkov *et al*., 2004).

Results and Discussion

Shown in Fig. 1 are NFP profiles of J1 and J2 type alphoids constructed using the tool RECON (Levitsky *et al*., 2001). NFP value at each point corresponds to the probability of positioning the center of nucleosome at this point. The maximal NFP values within J2 type monomers, interpreted as the signals for translational nucleosome positioning, are observed at three positions, namely, 378, 414, and 449. The type J1 monomers contain four–five such signals. This result complies well with the data of experimental studies on nucleosome positioning in the absence of CENP-B (Yoda *et al*., 1998).

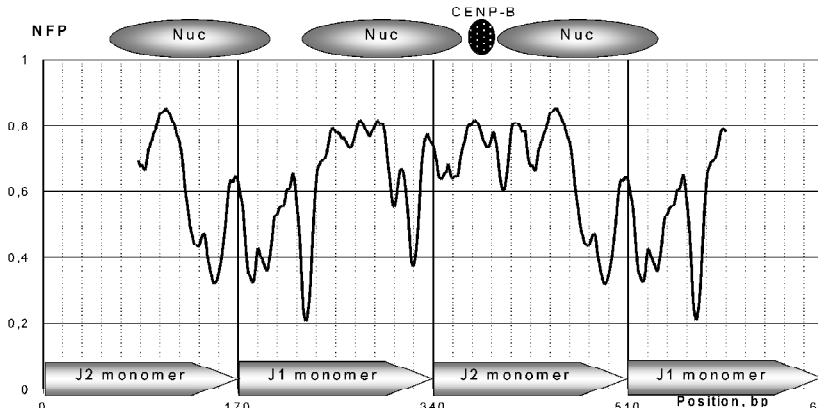


Fig. 1. NFP profile along J2 and J1 type alpha satellite sequences. The layout of nucleosome positioning within each monomer according to experimental data (Yoda *et al*., 1998; Ando *et al*., 2002) is shown above as well as the location of CENP-B protein over the B boxes in J2 type monomers.

Analysis of the sample of J1 and J2 type alpha satellites using the tool SITECON (Oshchepkov *et al*., 2004) detected existence of several blocks of conservative properties, in particular, the property Wedge, which is a geometric sum of the angles Roll and Tilt and characterizes a total curvature of a free B DNA (Ulanovsky, Trifonov, 1987). The property Wedge illustrates well the distribution pattern of conservative properties along the J1–J2 monomers.

The block of Wedge conservation in the J2 type monomer is localized to the region of B box, reflecting a high conservation of its context as a site for binding CENP-B protein (Yoda *et al*., 1998; Ando *et al*., 2002). However, the Wedge profile is insufficiently coordinated and hence, not additive. This means that the intrinsic curvature of the B box free of CENP-B protein is insignificant. Indeed, X ray structure analysis data demonstrate that B box as a DNA fragment with a length of 21 bp is bent to 60° only in the complex with CENP-B protein (Tanaka *et al*., 2001).

In the sample of J1 type monomers, the region 36–52, corresponding to the B box region in J2 type monomers, also displays a conservation of the property Wedge. However, unlike the genuine B box, increased Wedge values within this region are observed only in a 5-bp fragment, corresponding to a half-turn of the DNA helix (Fig. 2, above). Presumably, this is connected with a prominent intrinsic curvature of DNA within this region and reflects the linker function of the region, which, as a rule, is located between two nucleosomes positioned by the complex B box–CENP-B (see the layout in Fig. 1).

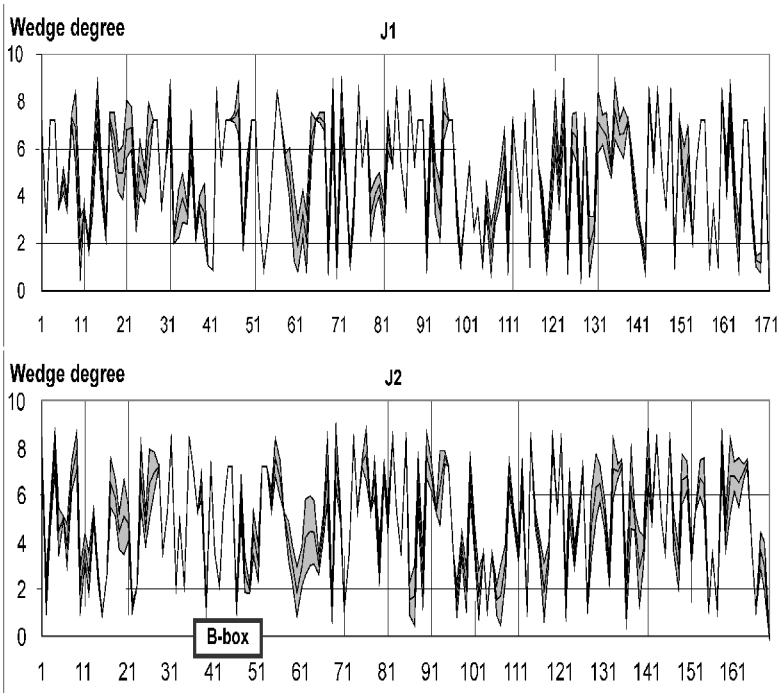


Fig. 2. Profile of the property Wedge along J1 type (above) and J2 type (below) monomers: firm line, the profile of mean values for the sample; gray area around the firm line, the range of standard deviation; the abscissa, nucleotide positions; and the ordinate, value of the property in degrees. Position of B box is shown.

Monomers of both types display conservation of the property Wedge over the region 97–105 as well. As is evident from Fig. 2, typical of this region are decreased values of the property in question, suggesting that DNA there is almost straight. It was demonstrated earlier that this particular conformation characterized the region of nucleosome dyad (Fitzgerald, Anderson, 1999); hence, this region may be assumed the center of nucleosome site, i.e., one of the preferable positions of the nucleosome dyad. Thus, the profile of the property Wedge along J1 and J2 type alpha satellites suggests a superposition of the context-dependent and CENP-B-dependent nucleosome positionings.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 03-04-48555-a and 03-07-96833); Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); Russian Federal Research Development Program «Research and Development in Priority Directions of Science and Technology» (contract No. 38/2004); and Program for Basic Research of the Presidium of the Russian Academy of Sciences (contract No. 10002-251/П-25/155-270/200404-082).

References

Ando S., Yang H., Nozaki N., Okazaki T., Yoda K. CENP-A, -B, and -C chromatin complex that contains the I-type alpha-satellite array constitutes the prekinetochore in HeLa cells // Mol. Cell. Biol. 2002. V. 22. P. 2229–2241.

Fitzgerald D.J., Anderson J.N. DNA structural and sequence determinants for nucleosome positioning // Gene Theor. Mol. Biol. 1999. V. 4. P. 349–362.

Gilbert N., Allan J. Distinctive higher-order chromatin structure at mammalian centromeres // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 11949–11954.

Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis // Bioinformatics. 2001. V. 17. P. 998–1010.

Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. SITECON—A tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition // Bioinformatics of genome regulation and structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London, 2004.

P. 93–102.

Tanaka Y., Nureki O., Kurumizaka H., Fukai S., Kawaguchi S., Ikuta M., Iwahara J., Okazaki T., Yokoyama S. Crystal structure of the CENP-B protein-DNA complex: the DNA-binding domains of CENP-B induce kinks in the CENP-B box DNA // EMBO J. 2001.V. 20. P. 6612–6618.

Ulanovsky L.E., Trifonov E.N. Estimation of wedge components in curved DNA // Nature. 1987. V. 326. P. 720–722.

Yoda K., Ando S., Okuda A., Kikuchi A., Okazaki T. *In vitro* assembly of the CENP-B/alpha-satellite DNA/core histone complex: CENP-B causes nucleosome positioning // Genes Cells. 1998.V. 3. P. 533–548.

BACTERIAL METAL RESISTANCE SYSTEMS REGULATED BY TRANSCRIPTION REGULATORS OF THE MERR FAMILY

***Kazakov A.E.*[1], *Kalinina O.V.*[2], *Permina E.A.*[3], *Gelfand M.S.*[1, 2, 3]**

[1] Institute for Information Transmission Problems, RAS, Moscow, Russia; [2] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; [3] State Scientific Center GosNIIGenetika, Moscow, Russia

***** Corresponding author: e-mail: kazakov@iitp.ru

Keywords: *metal resistance, transcription regulation, comparative genomics, MerR family*

Summary

Motivation: Understanding of the metal metabolism in bacteria is inseparable from investigation of metal resistance systems. One class of such systems is regulated by the MerR family of transcription regulators.

Results: We selected metal-sensing regulators belonging to the MerR family and explored genomic loci where these regulators were located. Based on these loci and several known sites, we identified candidate regulatory sites for MerR family proteins in the regulator loci and (in several cases) elsewhere in the analyzed genomes.

Introduction

Some metals, such as iron, copper, manganese, etc. are micronutrients used in the redox process, regulation of the osmotic pressure and as enzyme components. Other metals are not essential. However, even essential metals such as zinc and copper are toxic at high concentrations. To protect themselves from dangerous environment, bacteria have different mechanisms of cell resistance to toxic metals that involve permeability barriers, intra- and extracellular sequestration, efflux pumps, enzymatic detoxification and reduction.

One class of transcription factors regulating the metal resistance in bacteria is the MerR family, named after the regulator of mercury resistance. MerR-family proteins regulate systems of mercury detoxification (MerR), resistance to zinc (ZntR), copper (CueR and HmrR), cadmium (CadR) (Brown *et al*., 2003). All known MerR family regulators bind to palindromic sequences located between the –35 and –10 promoter boxes. In this study we analyzed systems of resistance to high concentration of copper, cadmium and zinc regulated by the members of the MerR family.

Materials and Methods

MerR-family proteins were retrieved from the SMART database (domain accession number SM00422) (http://smart.embl-heidelberg.de/). Multiple sequence alignments were done using the CLUSTALX program (Thompson *et al*., 1997). Phylogenetic trees were constructed using the program PROML from the PHYLIP package (the maximum likelihood method) (Felsenstein, 1996). A simple iterative procedure implemented in the software package GenomeExplorer was used to construct a profile from a set of upstream gene fragments and to search for possible regulatory sites in genomic sequences (Mironov *et al*., 2000).The positional nucleotide weights in these profiles were defined as (Mironov *et al*., 1999):

$$W(b,k) = \log [N(b,k) + 0.5] - 0.25 \sum_{i=A,C,G,T} \log [N(i,k) + 0.5], \quad (1)$$

where $N(b,k)$ denoted the count of nucleotide b at position k. The score of a L-mer candidate site was calculated as the sum of the respective positional nucleotide weights:

$$Z(b_1 \ldots b_L) = \sum_{k=1 \ldots L} W(b_k, k) \quad (2)$$

Results and Discussion

109 out of 503 MerR family members were selected as metal-sensing based on the presence of at least two out of three conserved cysteine residues required for the cation binding (Brown *et al*., 2003). The selected regulators were re-aligned and a phylogenetic tree was constructed. The branches containing known regulators CadR, ZntR, CueR, HmrR, MerR were identified on the tree, while several branches contained no regulators with known specificity.

In many cases, loci, harboring metal-sensing regulators are divergons containing two genes: one of them encoding a MerR-family transcriptional regulator and the other encoding a possible transporter. Some of these transporters were experimentally verified earlier (for a review see Brown *et al*., 2003). In other cases, no potential regulated transporters were found in the regulator neighborhood.

Most of the transporters belong to the P-type ATPase superfamily, TC# 3.A.3 according to the TC-DB Transport Classification Database (http://tcdb.ucsd.edu/tcdb/). Almost all members of this superfamily catalyze cation uptake and/or efflux driven by the ATP hydrolysis. On the phylogenetic tree of the superfamily, all transporters co-located with MerR-family transcriptional regulators cluster on neighboring branches, and transporters of the same presumed specificity belong to the same branches. However, in several cases putative transporters exhibit no homology to the P-type ATPase superfamily, nor to any other family in the TC-DB database. All these genes are preceded by strong CadR sites. Thus, the cadmium efflux may be facilitated by two non-orthologous protein families.

After that, we investigated the identified metal resistance loci for known or candidate regulatory sites. There are seven known regulatory sites confirmed experimentally for CueR, HmrR, CadR and ZntR (two sites per regulator, except the last one, for which one site is known) (Brown *et al*., 2003; Brocklehurst *et al*., 2003). The standard mechanism of regulation by MerR is well-known: it involves protein-protein interactions between MerR and RNA polymerase, accomplished by protein-induced DNA distortions (Brown *et al*., 2003), and requires precise location of a palindromic regulator binding site in relation to the -35 box of the promoter with an unusually long spacer (19 or 20 bp) between the -35 and -10 promoter boxes. Thus, we searched upstream regions of genes that could be regulated for palindromic sequences located in 19 or 20 bp spacers of candidate promoters.

The typical locus structures and consensus sequences of the identified signals are listed in Table. Briefly, for CadR and its homologs the typical locus structure is a divergon including a regulator and a transporter. In the case of the CueR regulator, the typical situation is a divergon or an operon including a regulator and an efflux pump, in several cases these genes are separated. ZntR orthologs and their regulated genes usually are not linked on the chromosome. In addition to genes enco-ding cation transporters, other potentially regulated genes were found. In *E. coli*, *S. typhimurium* and *Y. pestis*, a gene encoding multicopper oxidase has a potential CueR regulatory site. Other probable CueR sites were found upstream of genes encoding putative copper chaperone (in *S. typhimurium* and *P. putida*) and cytochrome c554 (in *V. vulnificus* and *V. parahaemolyticus*). In *Photorhabdus luminescens*, a potential ZntR binding site was found upstream of the gene *PLU4679* encoding a homolog of multidrug efflux proteins.

This analysis identified new candidate members of the metal-resistance regulons and binding signals of the MerR-family transcriptional factors. Clearly, these predictions should be verified in experiment, but they should facilitate the direct experimental approaches.

Acknowledgements

This study was partially supported by grants from HHMI (55000309), the Fund for Support of the Russian Science, and the Programs “Molecular and Cellular Biology” and “Origin and Evolution of the Biosphere” of RAS. We are grateful to Inna Dubchak for the suggestion to analyze the MerR family.

Table. Locus structures and regulatory site consensi for metal-sensing regulators of the MerR-family

Name	Genomes	Locus	Site
CadR	*Pseudomonas syringae,* *Pseudomonas putida,* ***Pseudomonas putida* plasmid pWW0, *Acinetobacter lwoffii* plasmid pKLH202,** *Pseudomonas aeruginosa,* ***Pasteurella multocida***	Regulator(*cadR*)/Transporter (*cadA*) divergon	ACCCTATAGNNNCTATGGGT
Proposed CadR	***Rhodopseudomonas palustris,* *Pseudomonas syringae***	Regulator/Transporter divergon	**ACCTGTAGNNNCTACAGGT**
	Mesorhizobium loti,* *Sinorhizobium meliloti,* *Agrobacterium tumefaciens,* *Brucella melitensis,* *Brucella suis	Regulator/Transporter divergon (in *S. meliloti* and *A. tumefaciens* regulator and transporter are separated)	**TCCTCTAGNNNCTAGAGGA**
ZntR	*Escherichia coli,* *Yersinia pestis,* *Salmonella typhi,* *Salmonella typhimurium,* ***Shewanella oneidensis,* *Photorhabdus luminescens***	Regulator and transporter are separated except in *S. odeniensis*, where they are constitute an operon	ACTCTGGAGTCGACTCCAGAGT
ZntR	***Vibrio parahaemolyticus,* *Vibrio vulnificus,* *Vibrio cholerae***	Regulator and transporter are separated	**ACCTTGGAGTCGACTCCAGGAT**
CueR	*Escherichia coli,* *Salmonella typhi,* *Salmonella typhimurium,* *Yersinia pestis,* ***Vibrio cholerae,* *Vibrio parahaemolyticus,* *Vibrio vulnificus***	Regulator/Transporter divergon (in *Salmonella* and *Yersinia*) or separated (all other cases). Genes encoding multicopper oxidase always separated	ACCTTCCCNNNNNGGGAAGGT
HmrR	*Agrobacterium tumefaciens,* ***Brucella suis,* *Mesorhizobium loti,*** *Rhizobium leguminosarum,* *Sinorhizobium meliloti,* *Sinorhizobium meliloti* symbiotic plasmids pSymA and pSymB, *Pseudomonas aeruginosa,* *Pseudomonas syringae,* *Pseudomonas putida,* ***Ralstonia solanacearum,* *Salmonella typhimurium***	Regulator and transporter usually constitutes an operon. In *P. aeruginosa*, *B. suis* and *M. loti* they are separated. Other co-regulated genes are separated from them	

Notes. Bold: genomes with MerR family regulators and signals identified in this study. Mercury resistance systems regulated by MerR are not shown.

References

Brown N.L., Stoyanov J.V., Kidd S.P., Hobman J.L. The MerR family of transcriptional regulators // FEMS Microbiol Rev. 2003. V. 27. P. 145–163.

Brocklehurst K.R., Megit S.J., Morby A.P. Characterisation of CadR from Pseudomonas aeruginosa: a Cd(II)-responsive MerR homologue // Biochem Biophys Res Commun. 2003. V. 308. P. 234–239.

Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods // Methods Enzymol. 1996. V. 266. P. 418–427.

Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes // Nucleic Acids Res. 1999. V. 27. P. 2981–2989.

Mironov A.A., Vinokurova N.P., Gelfand M.S. Software for analyzing bacterial genomes // Mol. Biol. 2000. V. 34. P. 222–231.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools // Nucleic Acids Res. 1997. V. 25. P. 4876–4882.

GENOME REVIEWS: INTEGRATED VIEWS OF COMPLETE GENOMES

Kersey P.J.*, Morris L., Faruque N., Kulikova T., Whitfield E., Apweiler R.

EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom
* Corresponding author: e-mail: pkersey@ebi.ac.uk

Keywords: *bioinformatics, databases, Genome Reviews, genomics, protein classification, proteome analysis, UniProt*

Summary

Motivation: Since the advent of whole genome sequencing in the mid 1990s, the sequences of over 170 organisms have been completely determined and deposited in the public repositories. However, submitters retain ownership of data in these repositories and in consequence annotation is often not standardized or updated. Therefore, we have launched Genome Reviews, a new project to make annotated genome sequence available in an EMBL-compatible format but with standardized, up-to-date annotation.

Results: Genome Reviews has been released with coverage of over 150 bacterial and archaeal genomes. Annotation in EMBL has been standardized and supplemented by data imported from curated resources such UniProt, InterPro and GOA. Evidence tags have been added into the format to enable the source of individual data items to be tracked.

Availability: Genome reviews are available at {{http://www.ebi.ac.uk/GenomeReviews/}}. Software for reading/writing Genome Reviews files is available from the authors on request.

Introduction

The International Nucleotide Sequence Database, a collaboration between the EMBL (Kulikova *et al*., 2004), Genbank (Benson *et al*., 2003) and DDBJ (Miyazaki *et al*., 2004) nucleotide sequence databases, is a repository for DNA and RNA sequence and annotation. Since the advent of whole genome sequencing in the mid 1990s, the sequences of over 170 organisms have been completely determined and deposited in the public repositories. The rate of deposition of such sequences is still increasing, with over 60 genomes sequenced between March 2003 and March 2004. Although the determination of mammalian genomes has attracted attention, the majority of sequenced organisms continue to be bacteria (140 sequenced organisms in March 2004) and archaea (18 sequenced organisms in March 2004).

Submitters retain ownership of data in the nucleotide sequence repositories and in consequence annotation is often not standardized or updated. Various problems can result from this: annotation may be poor; information may be represented irregularly; and as much annotation for predicted genes is inferred by similarity from other sequences, it becomes out-of-date as new sequences are annotated. Additionally, the integration of theoretical annotation inferred from sequence may not be integrated with data from laboratory experiments. Many of these issues are addressed in curated databases such as UniProt/Swiss-Prot; but annotation improvements/updates implemented in these resources cannot be incorporated in the archive record.

As an illustration of this problem, there are currently 4356 entries in UniProt that together define a non-redundant proteome set for the Gram-negative bacterium *Escherichia coli K-12*, representing the products of 4396 genes. Of these, some 1045 proteins (24 %) have been assigned a protein sequence by UniProt curators other than that originally predicted in the EMBL entry derived from the original genome submission (Blatter *et al*., 1997), which was last updated in 1998.

The NCBI RefSeq project (Pruitt *et al*., 2003) aims at providing compatible records for DNA, RNA and protein sequence for all complete genomes. RefSeq entries for bacteria frequently contain more regular annotation than the primary submissions with, for example, a standardized representation of gene names. However, little additional information is imported into RefSeq entries. Additionally, there is limited compatibility with external resources (for example, CDS features in RefSeq are not supplied with the common identifiers used by EMBL, Genbank and DDBJ, but instead use their own identifier space).

Therefore, we have launched Genome Reviews, a new project to make annotated genome sequence available in an EMBL-compatible format but with standardized, up-to-date annotation, derived from curated data sources such as UniProt, InterPro and GOA.

Methods and Algorithms

EMBL entries describe (nucleotide) sequences, features (annotated regions of sequence) and feature qualifiers (individual annotations attached to a feature). Additionally, there is also some annotation that is attached to the database entry itself as opposed to the sequence (for example, the database accession number). The 'CDS' (CoDing Sequence) feature is used to identify sub-sequences within the overall sequence that correspond with the sequence of nucleic acids in a protein (proteins are the most widely annotated biological entity); the '/db_xref' feature contains cross-references to entries in other databases. The '/protein_id' qualifier uniquely identifies each CDS.

There are therefore 3 ways in which an entry in another database can be identified as referring to the same biological entity as a given EMBL (CDS) feature: if the EMBL feature cross-references that entry, if that entry cross-references the EMBL feature, or if an entry in a third database cross references both other entries. Through tracking identifiers between databases, additional annotation belonging to a feature can be identified. UniProt (Apweiler *et al*., 2004), a well-annotated protein knowledgebase in which redundant submissions are merged, is a particularly useful hub database for retrieving annotation and cross-links to further resources. For each type of annotation, a particular preferred source is nominated; and annotation of that type from that is imported into corresponding features in Genome Reviews either as a supplement to or a replacement for the original annotation. Overlapping sets of annotations (e.g., gene names derived from different sources) are case-standardized and merged.

The annotation attached to other types of features (for example, non-coding RNAs) has also been standardized, and redundant or rarely used features and feature qualifiers removed. In addition, new features have been added (for examples, mature peptides produced after cleavage) by propagating features on protein sequences described in UniProt back onto the corresponding DNA. Sequences are first compared to ensure that the co-ordinate systems used for features are compatible.

Implementation and Results

Software for producing Genome Reviews has been implemented using the Java programming language (Java 2 Standard Edition v1.4.2), in particular utilizing tools provided by the BioJava project ({{http://www.biojava.org}}). A persistence layer has been implemented in a relational database management system (Oracle 8i Enterprise Edition release 8.1.7). This database, used to build each release, contains manually curated information used in making each entry (for example, corrected literature citations and taxonomy) and a representation of the relationships between genes, transcripts and proteins for all complete genomes.

Release 0.4 of Genome Reviews, made publicly available on 29th March 2004, contained files describing 256 chromosomes and plasmids from 153 bacterial and archaeal species. Some details of the change in the quantity of annotation are shown in the Table.

	Original EMBL entries	Genome Reviews Entries
Number of feature types	29	8
Number of qualifier types	42	26
Number of feature qualifiers	4132463	6165606
Number of external databases cross-referenced	9	18
Number of 'mat_peptide' features	0	3613
Number of '/db_xref' qualifiers	651012	2527269
Number of '/locus_tag' qualifiers	207477	459327
Number of evidence tags	0	4422463

The Table shows how in Genome Reviews we have standardized annotation, reducing the number of features and feature qualifiers actually in use, but increasing the total quantity of annotation. The creation of additional features, and the addition of extra qualifiers to existing features, is illustrated for the 'mat_peptide' feature type and the '/db_xref' feature qualifier respectively. The increased use of the '/locus_tag' feature qualifier indicates the proper use of this qualifier to indicate the systematic gene names (such information is found variously, if at all, in the primary EMBL submissions).

The introduction of evidence tags into Genome Reviews entries represents the most significant change to EMBL flat file format in Genome Reviews. Evidence tags provide a simple record of the source of each piece of information in a Genome Reviews entry. Tags are applied to the feature qualifiers (indicating why the information contained in a qualifier has been applied to the corresponding feature). If a tag is applied to the '/evidence' qualifier, this indicates why the feature has been attached to the sequence.

Discussion

In databases such as Ensembl (Birney *et al*., 2004), consisting wholly of predictions on a genomic sequence, a set of complete theoretical genes, transcripts and proteins are provided. In the case of many bacterial genomes, however, it is known that many of these theoretical gene predictions are wrong (Skovgaard *et al*., 2001), and annotation is not regularly updated. In Genome Reviews, we are attempting to standardize and update annotation in line with experimental results.

In addition to tracking identifiers as described, it is also possible to co-identify equivalent biological entities described in different databases through their common co-ordinates on a shared reference sequence; to identify features unannotated in the genome; or identify disagreements in sequence between entries known (through use of common identifiers) to represent the same entity. In future releases of Genome Reviews, we will add new feature qualifiers to describe disagreement at the sequence level, to support the consistent representation of genomic DNA and experimentally verified protein sequence.

Acknowledgements

This work has been funded by the award of grant number QLRI-CT-2001000015 from the European Union under the RTD program "Quality of Life and Management of Living Resources".

References

Apweiler R. *et al*. UniProt: the universal protein resource // Nucleic Acids Res. 2004. V. 32. D115–D119.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. Genbank // Nucleic Acids Res. 2003. V. 31. P. 23–27.

Birney E. *et al*. Ensembl 2004 // Nucleic Acids Res. 2004. V. 32. D468–D470.

Blatter F.R. *et al*. The complete genome sequence of *Escherichia coli* K-12 // Science. 1997. V. 277. P. 1453–1474.

Kulikova T. *et al*. The EMBL nucleotide sequence database // Nucleic Acids Res. 2004. V. 32. D27–D30.
Miyazaki S., Sugawara H., Ikeo K., Gojobori T., Tateno Y. DDBJ in the stream of various biological data // Nucleic Acids Res. 2004. V. 32. D31–D34.
Pruitt K.D., Tatusova T., Maglott D.R. NCBI reference sequence project: update and current status // Nucleic Acids Res. 2003. V. 31. P. 34–37.
Skovgaard M., Jensen L.J., Brunak S., Ussery D., Krogh A. On the total number of genes and their length distribution in complete microbial genomes // Trends Genet. 2001. V. 17. P. 425–428.

COMPARISON OF THE STRUCTURES OF *IN VITRO* SELECTED AND NATURAL BINDING SITES OF TRANSCRIPTION FACTORS

Khlebodarova T.M.*, Podkolodnaya O.A., Ananko E.A., Ignatieva E.V.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: tamara@bionet.nsc.ru

Keywords: *databases, transcription factor, binding sites, transcription regulation*

Resume

The information compiled in TRRD_ARTSITE database was used to compare the structures natural and *in vitro* selected binding sites for transcription factors USF1, SP1, YY1, RAR/RXR, and E2F/DP1. The structures of the natural and *in vitro* selected binding sites for each transcription factor analyzed were found similar, suggesting that at least for the factors in question the structures of binding sites correlated with the affinities of the corresponding factors. Insignificant differences in frequencies of certain nucleotides detected reflect the general trend, namely, a higher occurrence of moderate affinity sites in the natural population of sequences compared with the sequences obtained *in vitro*.

Introduction

Recently, development of new technologies, in particular, SELEX technologies, yielded numerous data on the structures of binding sites for various transcription factors, both eukaryotic and prokaryotic. However, the questions on whether these data reflect the genuine structures of natural binding sites and what are the potential of applying these data to search for and prediction of natural sites are yet to be answered. The opinions on this issue are inconsistent and ambiguous and unfavorable for at least several prokaryotic transcription factors (Robison *et al*., 1998). Thus, it is no wonder that having a considerable volume of information on the structures of natural binding sites complied with the database TRRD (Kolchanov *et al*., 2002). We developed the database TRRD_ARTSITE, whose contents allowed us to perform comparative studies of the structures of natural and *in vitro* selected binding sites for the transcription factors USF1, SP1, YY1, RAR/RXR, and E2F/DP1.

Comparison of in vitro selected and natural binding sites for USF1, RAR/RXR, and SP1 transcription factors. Thus, what are the potential of using the data obtained as a result of *in vitro* selection for recognition of *in vivo* sites in genomes of various organisms? To clarify this issue, we compared the matrices constructed using the sequences selected *in vitro* with the matrices constructed using the sequences extracted from TRRD. Table 1 shows the data obtained upon such comparison for three transcription factors having different DNA-binding domains and different types of binding to DNA. The matrices constructed using different sources are nonetheless virtually similar with reference to most frequently occurring nucleotides within the detected cores for USF1 and SP1 transcription factors and differ inessentially in significant nucleotides at positions –2, –3, and +4 in the case of RAR/RXR heterodimer (Table 1). In the last case, G occurs most frequently in the natural sites at position –2, whereas G or A in the in vitro selected sequences; A and T are met with equal probabilities at position –3 in natural sites, whereas A is typical of in vitro selected sites; and A is most frequent at position +4 in natural sites, whereas A and G are equiprobable in the in vitro selected sequences. Thus, at least one of the significant nucleotides is necessarily present at all the three positions in both matrices, suggesting that the distinctions detected are not crucial and may stem from a moderate size of one of the matrices. Note in this connection that a similar fact played no negative role when comparing *in vitro* selected and natural binding sites for SP1 transcription

factor. The matrix for this factor constructed basing on natural sequences and containing 244 functional sites did not differ in significant nucleotides from the matrix containing only 11 sequences detected by EMSA. We believe that these data indicate that the functionality of SP1 binding site correlates directly with the affinity of this transcription factor for the site in question.

Table 1. Matrices describing the binding sites for USF1, SP1, and RAR/RXR transcription factors: N, natural and S, *in vitro* selected

Factor Binding type DNA-binding domain	Nucleotide	Nucleotide position									
		-5	-4	-3	-2	-1	0	+1	+2	+3	+4
USF1 Homodimer bHLH domain	NA	**19**	3	0	**40**	0	3	0	0	**17**	3
	NC	9	**21**	**44**	1	**32**	7	0	0	7	**21**
	NT	1	**13**	0	2	3	0	**44**	0	5	7
	NG	**15**	7	0	1	9	**34**	0	**44**	**15**	**13**
	Consensus	R	Y	C	A	C	G	T	G	R	S
	SA	**8**	1	1	**30**	1	1	0	0	**9**	3
	SC	2	**16**	**30**	1	**28**	1	2	0	3	**12**
	ST	0	**11**	0	0	2	0	**28**	0	5	**13**
	SG	**7**	2	0	0	0	**29**	1	**31**	**14**	1
	Consensus	R	Y	C	A	C	G	T	G	R	Y
	(AS00239 - accession number ARTSITE_DB)										
SP1 Monomer Zinc fingers C2H2 type	NA		41	42	3	5	55	5	8	41	27
	NC		18	10	9	7	**144**	8	4	19	39
	NT		39	6	8	5	34	14	21	14	16
	NG		**147**	**186**	**224**	**227**	11	**216**	**210**	**169**	**161**
	Consensus		G	G	G	G	C	G	G	G	G
	SA		2	3	0	0	0	3	0	1	1
	SC		1	1	0	0	**8**	0	1	0	2
	ST		2	1	0	0	1	2	3	2	1
	SG		**6**	**6**	**11**	**11**	2	**6**	**7**	**8**	**7**
	Consensus		G	G	G	G	C	G	G	G	G
	(AS00292 - accession number ARTSITE_DB)										
RAR/RXR Heterodimer Nuclear receptor zinc fingers C4-type	NA		6	**13**	6	**12**	2	2	2	1	**29**
	NC		**18**	7	5	3	0	0	3	**28**	2
	NT		2	**10**	5	2	0	7	**27**	1	1
	NG		8	4	**18**	**17**	**32**	**25**	2	4	2
	Consensus		C	W	G	R	G	G	T	C	A
	SA		**9**	**13**	**7**	**9**	1	0	1	1	**6**
	SC		**4**	2	1	0	0	0	1	**11**	2
	ST		0	1	1	2	0	2	**11**	0	2
	SG		**3**	0	**7**	**5**	**15**	**14**	3	4	**6**
	Consensus		V	A	R	R	G	G	T	C	R
	(AS00305 - accession number ARTSITE_DB)										

Comparison of in vitro selected and natural binding sites for E2F/DP1 and YY1 transcription factors and assessment of their functionality. At present, the resource of our database allows for assessment of probable functionality of the sites detected basing on comparison of natural and artificially selected sequences for 17 transcription factors. For example, in Table 2 there are the matrices describing the binding sites for two transcription factors, E2F/DP1 and YY1. Comparison of the matrices constructed basing on the natural functional binding sites for YY1 transcription factor and the *in vitro* selected sequences suggests that the flanking nucleotides do not play a significant role in the function of these sites. Moreover, data on the affinity of the sequences detected in *in vitro* experiments and displaying significant flanking nucleotides (C, G, and G at

positions–4, –3, and +5, respectively) demonstrate that these nucleotides also have no effect on the level of binding, as they are met in all the types of sequences—with high, medium, and low affinities (Table 3). Analysis of these sequences demonstrates that the detected nucleotides (at positions –4 and –3) are contained in the primer used for selection and thus, were selected randomly. As for the G nucleotide at position +5, its appearance is not so evident. It cannot be excluded that the last nucleotide is necessary for the site function; however, a small sample of the natural sites prevented from detection of its significance. Nonetheless, note a trend of increase in its occurrence in the natural population and that its frequencies in natural and *in vitro* selected populations are rather similar: 41 and 53 %, respectively. Moreover, analysis of these matrices and the data listed in Table 2 allowed us to detect within the site the nucleotides responsible for the affinity of the factor for DNA. All the four nucleotides are met at position +2 with equal probabilities; however, only the presence of nucleotide A decreases drastically the affinity of the site (Table 3).

Table 2. Matrices describing the binding sites for E2F/DP1 and YY1 transcription factors: N, natural and S, *in vitro* selected

Factor	Nucleotide	Nucleotide position											
		-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5	+6
YY1	NA		4	5	2	4	**17**	2	4	3	3	4	
	NC		9	7	**19**	**17**	1	0	7	3	5	6	
	NT		4	5	1	1	4	**19**	9	**14**	**12**	3	
	NG		5	5	0	0	0	1	2	2	2	9	
	Consensus		N	N	C	C	A	T	N	T	T	N	
	SA		2	0	0	0	**56**	0	9	4	0	9	
	SC		**39**	0	**56**	**56**	0	0	12	1	2	6	
	ST		0	15	0	0	0	**56**	25	**50**	**50**	9	
	SG		15	**41**	0	0	0	0	10	1	3	**30**	
	Consensus		C	G	C	C	A	T	N	T	T	G	
		(AS00001 - accession number ARTSITE_DB)											
E2F/DP1	NA	3	4	0	1	0	0	0	0	2	**21**	**17**	**16**
	NC	1	0	0	**22**	**12**	**36**	2	**26**	**16**	7	4	3
	NT	**29**	**32**	**37**	1	0	0	0	0	2	4	**11**	**11**
	NG	5	2	1	**14**	**26**	2	**36**	**12**	**18**	6	6	8
	Consensus	T	T	T	S	S	C	G	S	S	A	W	W
	SA	4	2	0	0	0	0	3	0	0	**9**	**5**	**5**
	SC	2	1	1	**17**	**14**	**24**	1	**19**	**18**	4	2	1
	ST	**18**	**22**	**24**	0	0	0	0	0	4	**10**	**16**	**17**
	SG	1	0	0	**8**	**11**	1	**21**	**6**	3	1	1	0
	Consensus	T	T	T	S	S	C	G	S	C	W	W	W
		(AS00123 - accession number ARTSITE_DB)											

Similar analysis of the binding site for E2F/DP1 transcription factor was made. The matrices from Table 2 exhibit very close structures, namely, stringently fixed C and G nucleotides at positions 0 and +1 and inessential differences in significant nucleotides at positions +3 and +4. In natural sites, nucleotide A occurs most frequently at position +4, whereas in the *in vitro* selected, T and A. As for position +3, C and G are equiprobable in the natural sites, whereas C nucleotide is most frequent in the *in vitro* selected sequences. Thus, at least one of the significant nucleotides is present at these positions in both matrices, making the distinctions detected not principal. Presumably, these distinctions are related to a higher frequency of moderate affinity sites in the population of natural sequences. Data shown in Tables 2 and 3 give grounds for such inference. Presumably, the presence of C nucleotide in the core sequence (CCCGCC) is characteristic of high affinity sites, whose occurrence rate is higher among the *in vitro* selected sequences; however,

G at position +1 is the most significant nucleotide in the site structure. Other conditions being equal, substitution of G with A decreases drastically the affinity of the factor for DNA, and only a high resolution of the method used allowed such sites to be detected among the *in vitro* selected sequences. A virtually complete absence of the natural sites with any other nucleotides except for G at this position also confirms the significance of this position.

Table 3. Binding sites for transcription factors YY1 and E2F/DP1 detected in in vitro experiments and their degree of affinity for the corresponding factors

TF	Sequence	Affinity
YY1	cgCCATTTTaag	High
	gtCCATTTTtgt	Medium
	atCCATCTTgac	Medium
	cgCCATGTTgcg	Medium
	cgCCATTTGccg	Medium
	cgCCAT**A**TTcct	Low
	cgCCAT**A**TTgtc	Low
	gtCCAT**A**TTgta	Low
	(AS00001: AC ARTSITE_DB)	
E2F1/ DP1	ttattTTTCCCGCCTTT	High
	tCTTCCCGCCTTAttc	High
	tgatTTTGGCGGGATTc	Medium
	ttGTTCCC**A**GCCACtc	Very low
	(AS00123: AC ARTSITE_DB)	

Thus, the above examples demonstrate that analysis and comparison of the matrices constructed basing on the *in vitro* selected and natural sites expand considerably our knowledge on the structure of the site as well as suggest applying more optimistically various methods for detecting sites in unstudied genes and assessing theoretically their functionality.

A comprehensive analysis of the structure of CNF/NF1 binding sites using *in vitro* selected and natural site sequences gave similar results. This allowed the authors to use the results obtained for developing a highly sensitive method for prediction and recognition of DNA-binding sites in eukaryotic genomes (Roulet *et al*., 2000, 2002).

Acknowledgements

Work was supported in part by the RFBR (03-07-90181-в, 03-04-48469-a), RAS 10.4, Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), SB RAS (Integration Projects No.119), NIH USA (No.2 R01-HG-01539-04A2), Russian Federal Research Development Program (contract No. 38/2004), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”.

References

Kolchanov N.A., Ignatieva E.V. *et al*. Transcription regulatory regions database (TRRD): its status in 2002 // Nucl. Acids Res. 2002. V. 30. P. 312–317.

Robison K., McGuire A.M., Church G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome // J. Mol. Biol. 1998. V. 284. P. 241–254.

Roulet E., Bucher P., Schneider R. *et al*. Experimental analysis and computer prediction of CTF/NFI transcription factor DNA binding sites // J. Mol. Biol. 2000. V. 297. P. 833–848.

Roulet E., Busso S., Camargo A.A. *et al*. High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites // Nat. Biotechnol. 2002. V. 20. P. 831–835.

TRRD_ARTSITE DATABASE: STRUCTURES OF TRANSCRIPTION FACTOR BINDING SITES

Khlebodarova T.M.*, Podkolodnaya O.A., Ananko E.A., Stepanenko I.L., Ignatieva E.V., Podkolodny N.L., Pozdnyakov M.A., Proscura A.L.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: tamara@bionet.nsc.ru

Keywords: *databases, transcription factor, binding sites*

Resume

The TRRD_ARTSITE_DB database was developed; the database compiles the information on the structures of eukaryotic transcription factor binding sites and/or their DNA-binding domains obtained from natural or *in vitro* selected sequences. Current release of the database comprises 395 matrices describing specific features of binding sites or their DNA-binding domains for over 200 transcription factors. The matrices were constructed basing on alignments of representative samples of transcription factor binding sites, totally containing over 10 thousand sequences.

Introduction

During the last decade, an ever increasing attention is paid to creation of databases in various branches of knowledge, including regulation of transcription both eukaryotic and prokaryotic (Kolchanov *et al.*, 2002; Lescot *et al.*, 2002; Praz *et al.*, 2002; Wingender *et al.*, 2001; Munch *et al.*, 2003; etc.). An example is TRRD—Transcription Regulatory Regions Database of eukaryotic genes, whose development was commenced at the Institute of Cytology and Genetics SB RAS over 10 years ago. By now, TRRD contains a considerable array of data on the structure of promoters of various eukaryotic genes (Kolchanov *et al.*, 2002). However, recently appearing numerous data on the structure of binding sites for various transcription factors, including eukaryotic transcription factors, which previously were incompatible with the TRRD principles and, correspondingly, were not included in this database, could nonetheless expand considerably our knowledge on structure of the sequences interacting with transcription factors. Here, we mean the data on the structure of binding sites for transcription factors and/or their DNA-binding domains obtained using SELEX technologies. Emergence of the databases, such as SELEX_DB and JASPAR (Ponomarenko *et al.*, 2000; Sandelin *et al.*, 2004), which are trying to systematize this type of information, confirms the importance of this knowledge. However, the resource of databases indicated is yet rather small—116 and 111 entries, respectively. This type of information is also compiled with TRANSFAC, one of the largest databases integrating the information pertaining to regulation of transcription (Matys *et al.*, 2003). Unfortunately, the most part of these data is beyond the public domain. Thus, it is natural that having a considerable volume of information on the structures of natural binding sites in the database TRRD, we followed the path of integrating the data obtained using SELEX technologies with the data already compiled in TRRD.

Description of the format of ARTSITE database for *in vitro* selected sequences

One entry of ARTSITE_DB corresponds to one selection experiment where in a matrix describing a binding site for a transcription factor or one of its domains, provided the factor interacts with DNA in a complex manner, is obtained. This format also allows for describing binding sites for heterodimeric proteins and intricate complexes of transcription factors. Description of such an entry is shown in Fig.; the entry comprises 36 fields; of them, 24 fields are obligatory for filling in.

```
ID   BS_ELK1_SRF
DT   11/06/03
AC   AS00321
CR   Khlebodarova T.M.
DR   SWISS-PROT; ELK1_HUMAN; P19419;
DR   SWISS-PROT; SRF_HUMAN; P11831;
XX
TF   ELK1; ETS-domain protein ELK-1
TS   Homo sapiens
MM   ternary complex with SRF
DB   Ets domain
XX
TF   SRF; Serum response factor
TS   Homo sapiens
MM   ternary complex with ELK-1
DB   MADS domain
XX
MX   5'-CAGGTCAGTTCAGCGTCTAGAGTCCTTATATGG-(N32)-GAGGCGAATTCGTG-3'
RS   3
NS   60
AG   in vitro DNA-binding site selection
AG   EMSA with in vitro translated proteins
AS   1.3-13;+;           ccgtgaggtaccacttCCGGAAtggcttaacg
     1.3-36;+;       cgaatattcaaatcccaaccCCGGAAacccgc
     1.3-32;-;            tcctgcgtggatttCCGGGCcttagcgacgag
CC   SRF-binding site is present at primer sequences
CC   Sequences was aligned around Ets motif
XX
MA   13  15  19   7   9   0   0  55  43  18   8  13
MG   13  12  17   8   3  60  60   3   3  28  10  13
MC   18  13   6  41  47   0   0   0   2   5  15  13
MT   14  19  18   4   1   0   0   2  11   7  25  19
CN    N   N   D   C   C   G   G   A   A   R   B   N
XX
WA   0.1  0.2  0.8   0   0   0   0  52  25 0.7   0  0.1
WC   3.8  1.1    0  36  51   0   0   0   0   0 2.0  1.1
WT   0.1  0.8  0.6   0   0   0   0   0   0   0 3.6  0.9
WG   0.9  0.6  2.5 0.1   0  96  96   0   0  13 0.2  1.0
CN    S    N    D    C   C   G   G   A   A   G   B   N
XX
NM   1.3-13;+;
SQ   gaattcgcctcccgtgaggtaccacttccggaatggcttaacgccatataaggactctaga
IP   ----------------------------gg------------------------------; ELK-1;
EF   med affinity
XX
AU   Treisman R., Marais R., Wynne J.
TI   Spatial flexibility in ternary complexes between SRF and its
TI   accessory proteins.
SO   EMBO J.
VL   11
IS   11
YR   1992
PG   4631-4640
ML   1425594
```

Fig. An example of the entry of ARTSITE_DB database describing sites for binding of the transcription factor ELK1 to DNA detected in *in vitro* experiments.

The field names and pattern of their filling are made maximally similar to the pattern of TRRD and detailed below. ID is the identifier, which includes short name of the factor; DT, data of entry creation; AC, accession number in ARTSITE_DB; CR, name of the annotator who created the entry; DR, references to SWISSPROT and TrEMBL databases; TF, short and full names of the transcription factor; TY, synonyms of the transcription TG, TU, and TC, organ, tissue, and

cell line wherefrom the transcription factor was isolated, in the case it is of endogenous origin; MM, the form of binding of the transcription factor to DNA; DB, name of the DNA-binding domain of the transcription factor described; MX, the synthetic template used for selection experiment; RS, number of selection rounds; NS, number of sequences binding this transcription factor and detected in the experiment; AG, brief description of the methods used for selecting sequences; AS, aligned sequences from the original publication; MA (WA), MG (WG), MC (WC), and MT (WT), frequencies (or weights) of nucleotides at a certain position within the sequence aligned with respect to most frequently met nucleotides; CN, consensus; NM, number or name of the sequence in this series of experiments and its direction relative to the template used for amplification; SQ, sequence; IP, the positions important for binding of the transcription factor to DNA; EF, affinity of the sequence for the transcription factor; AU, authors of the publication annotated; TI, title of the publication; SO, brief name of the journal; VL, IS, PG, and YR, volume, number, pages, and year of issue; and ML, accession number in PubMed database. The field CC contains various textual comments on the specific features of experiment performed and construction of the weight matrix for the binding site described, which cannot be input in the format developed, but are important from the annotator's standpoint for a correct understanding of the data presented.

Specificity of the ARTSITE database format for natural sites

This database contains two types of entries describing matrices for natural binding sites. The first type is the entries obtained by selection of cloned genomic DNA fragments, and in this case, the format is virtually similar to that described above except that the field MA is not filled in. The second type of entries was obtained by aligning binding sites extracted from TRRD. In this case, each entry is obtained by annotating a large number of original publications describing transcription factors of various origins; therefore, the fields TS, AU, TI, SO, VL, IS, PG, YR, and ML, described above are not filled in. The fields related to description of the type of selection experiment are also left blank, while the field NM contains the accession number of the site in TRRD.

The content of ARTSITE database

The ARTSITE database is a natural extension of the database TRRD. The first release of the former database contains 395 matrices describing the binding sites for over 200 transcription factors and their DNA-binding domains. Of them, 345 matrices were constructed basing on alignment of 8940 sequences detected using various variants for selecting transcription factor binding sites described in 173 original publications and 50 matrices describing natural, functional binding sites for 44 transcription factors. The latter 50 matrices were constructed basing on alignment of 1409 sequences extracted from TRRD. The frequency matrices were used to construct the corresponding weight matrices. An example of the weight matrix describing the structure of the Ets domain of ELK1 transcription factor binding site is shown in Fig. (see the fields WA, WT, WC, and WG). The consensus obtained from the weight matrix differs insignificantly fRom the consensus derived from the frequency matrix; however, the significance of individual nucleotides is more distinctly evident in the weight matrix. The database content with reference to the structure of DNA-binding domains is shown in Table.

Thus, this work describes the format and content of the database ARTSITE. This is a database on the structures of both natural and *in vitro* selected eukaryotic transcription factor binding sites. It is a natural extension of the database TRRD and comprises now almost 400 matrices describing the structures of over 10 thousand binding sites for more than 200 transcription factors and/or their DNA-binding domains. Note in conclusion that although a relatively small volume yet limits the possibilities of ARTSITE_DB, its content provides search for and recognition of

transcription factor binding sites and their DNA-binding domains in various eukaryotic genomes. A confirmation is the development of several resources (http://wwwmgs2.bionet.nsc.ru:8080/tfbs_analyzer/; http://wwwmgs2.bionet.nsc.ru/cgi-bin/mgs/sitecon/sitecon.pl?stage=0 and others) basing on ARTSITE_DB, which allow for performing this type of search using various recognition methods.

Table. The ARTSITE database content with reference to the structure of transcription factor DNA-binding domains

DNA-binding domain	Number of matrices constructed using the sequences selected in vitro	Number of matrices constructed using natural sites
Basic domain	93	16
CUT repeat	11	-
Ets-domain	19	-
Homeodomain	48	6
HMG box	13	1
MADS	18	1
Myb domain	18	-
Nuclear receptor type	25	7
Paired domain	6	1
POU domain	12	-
Zinc finger	86	10
Others	14	8
Total	345	50

Acknowledgements

Work was supported in part by the RFBR (03-07-90181-в, 03-04-48469-a), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), SB RAS (Integration Projects No.119), NIH USA (No.2 R01-HG-01539-04A2), Russian Federal Research Development Program (contract No. 38/2004), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”.

References

Kolchanov N.A., Ignatieva E.V. *et al*. Transcription regulatory regions database (TRRD): its status in 2002 // Nucl. Acids Res. 2002. V. 30. P. 312–317.

Lescot M., Dehais P., Thijs G. *et al*. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences // Nucl. Acids Res. 2002. V. 30. P. 325–327.

Munch R., Hiller K., Barg H. *et al*. PRODORIC: prokaryotic database of gene regulation // Nucl. Acids Res. 2003. V. 31. P. 266–269.

Matys V., Fricke E., Geffers R. *et al*. TRANSFAC: transcriptional regulation, from patterns to profiles // Nucl. Acids Res. 2003. V. 31. P. 374–378.

Ponomarenko J.V., Orlova G.V., Ponomarenko M.P. *et al*. SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation // Nucl. Acids Res. 2000. V. 28. P. 205–208.

Praz V., Perier R., Bonnard C., Bucher P. The eukaryotic promoter database, EPD: new entry types and links to gene expression data // Nucl. Acids Res. 2002. V. 30. P. 322–324.

Sandelin A., Alkema W. *et al*. JASPAR: an open-access database for eukaryotic transcription factor binding profiles // Nucl. Acids Res. 2004. V. 32. D91–94.

Wingender E., Chen X., Fricke E. *et al*. The TRANSFAC system on gene expression regulation // Nucl. Acids Res. 2001. V. 29. P. 281–283.

TRANSLATIONAL POLYMORPHISM AS A POTENTIAL SOURCE OF EUKARYOTIC PROTEINS VARIETY

Kochetov A.V. [1*], ***Sarai A.***[2], ***Kolchanov N.A.***[1]

[1]Institute of Cytology and Genetics, Novosibirsk, Russia; [2] Kyushu Institute of Technology, Dept. Biochemical Engineering and Science, Iizuka, 820-8502 Japan
* Corresponding author: e-mail: ak@bionet.nsc.ru

Keywords: *mRNA, translation, proteome, organelle*

Resume

Motivation: According to scanning model, 40S ribosomal subunits can either initiate translation at start AUG codon in suboptimal context or scanthrough and initiate translation at downstream AUG(s). Functional significance of such a protein translational polymorphism is still unknown.

Results: We compared predicted subcellular localizations of annotated *Arabidopsis thaliana* proteins and their potential N-terminally truncated forms started from the nearest downstream in-frame AUG codons. It was found that localizations of full and N-truncated proteins differ in many cases: *ca*. 4% of *Arabidopsis thaliana* genes tested could produce additional protein forms with new targeting signals. It is likely that the in-frame downstream AUGs may be frequently utilized to synthesize proteins possessing new functional properties and such a translational polymorphism may serve as an important source of cellular and organelle proteomes.

Availability: Detailed description of *Arabidopsis* mRNAs potentially producing novel functional protein forms is available by request.

Introduction

Arabidopsis thaliana genome was assumed to contain ca. 25000 protein coding genes. The number of proteins actually formed may be considerably higher because of alternative splicing. Another possible source of new protein forms is translational polymorphism where several AUG codons within mRNAs may serve as alternative translation start sites (TSSs) to produce proteins with overlapping sequences and displaying different properties. Despite many examples where such mRNAs were described (*e.g.*, Watanabe *et al*., 2001), the contribution of translational polymorphism to proteome complexity was not evaluated.

According to the scanning model, 40S ribosomal subunits can either initiate translation at start AUG codon in suboptimal context or miss it and initiate translation at downstream AUG(s). The initiation/scanthrough ratio depends on both the translation start site context and the features of downstream mRNA fragment (Lukaszewicz *et al*., 2000; Kozak, 2002).

It was found that a part of cellular mRNAs with start AUG codon lying in suboptimal context is relatively large as well as a part of mRNAs with AUG-containing 5' untranslated regions (5'-UTRs) (Rogozin et al., 2001). It is likely that at least some mRNAs with suboptimal start codon context may produce two or more functional polypeptides. To test this assumption we isolated a sample of Arabidopsis thaliana mRNAs and compared predicted subcellular localizations of polypeptides started from annotated suboptimal TSS and the nearest downstream in-frame AUG codons.

Methods

EMBL cDNA entries were obtained at http://srs.ebi.ac.uk/ using the following search fields and terms: "Organism" – Arabidopsis thaliana; "Molecule" – RNA; "FtKey" – CDS; "Description" –

complete CDS. 9531 non-redundant sequences contained both complete coding parts and 5'-UTRs longer than 20 nucleotides. Subcellular localizations of proteins were evaluated by TargetP prediction progam (Emanuelsson et al., 2000) used with default parameters.

Results and Discussion

Analysis of nucleotide frequencies in AUG context positions showed that 24% of *Arabidopsis* mRNAs contain suboptimal start codon context (*e.g.*, they contained pyrimidines in position –3). We assumed that pyrimidines at position –3 will result in alternative translation from downstream AUG although the recognition/scanthrough ratio may vary and depend on the other context positions (-2,-1,+4,+5) as well as some other mRNA features (Kozak, 2002). mRNAs with G in position –3 were excluded because this nucleotide influences AUG "strength" in a tissue-specific manner (Lukaszewicz *et al*., 2000). A sample of *Arabidopsis thaliana* mRNAs with suboptimal TSS was analyzed. 130-amino acids long N-terminal protein fragments were used. Two samples were generated: either started from annotated TSS or from the nearest downstream in-frame AUG codons. In total, subcellular localization of 1819 full and N-truncated proteins were compared.

The results of prediction are shown in Table. One can see that N-truncated forms of many secreted polypeptides lose their targets. It was expected since N-truncated polypeptides could lose their secretory leader peptides. However, 12.2 % of N-truncated proteins acquire sorting signals de novo and 5.7% change their predicted subcellular locations (mitochodria, chloroplast or secretory pathway; detailed description is available by request). It may mean that a substantial part of Arabidopsis thaliana mRNAs produce two (or more?) proteins each with different subcellular localization due to translational polymorphism.

It was proposed that *Arabidopsis thaliana* chloroplast and mitochondria proteomes contain ca. 3100 different proteins each (Leister, 2003) but only a small number of proteome constituents were identified (Peltier *et al*., 2002; Kumar *et al*., 2002). Translational polymorphism allows to generate two or more protein forms. It might represent an appropriate way to address proteins of the same function to different locations or generate protein forms with different functions (e.g., Watanabe *et al*., 2001). Note, that the ratio of full- and N-truncated protein forms may be tightly regulated through adjustment of the start codons contexts to control initiation/scanthrough ratio (Kozak, 2002) that may provide a unique mechanism of efficient expression control.

Table. Subcellular localization of full and N-truncated proteins (%) predicted with *TargetP* program (Emanuelsson *et al*., 2000)*

Full		N-truncated			
Location	size of fraction	mTP	cTP	SP	others
mTP	11	2	0.9	1	7.1
cTP	20.3	1.5	6.8	0.5	11.5
SP	12.3	1.2	0.6	4.7	5.8
Others	56.4	5.4	3.2	3.6	44.2
Total	100	10.1	11.5	9.8	68.6

*mTP, mitochondria targeting peptide; cTP, chloroplast targeting peptide; SP, secretory peptide.

We proposed that *Arabidopsis* genes with suboptimal start codon context could encode additional protein forms with new functions and demonstrated that full and potential N-truncated protein forms frequently differ in subcellular localization. Indeed, this evaluation of impact of translational polymorphism and a possible functional role of N-truncated *Arabidopsis thaliana* proteins was very rough because of only the difference in subcellular sorting of protein forms was taken into account. It is clear that they may differ in many other features and functional properties. It is likely that at least some N-truncated proteins are synthesized in plant. Thus, translational polymorphism may be considered as an important source of cellular and subcellular proteomes and should be taken

into account in further investigations. Recent evaluation showed that many eukaryotic genes yield transcript(s) that translate into several, and often very numerous families of polypeptide species (Kettman *et al*., 2002). Further experimental and theoretical estimations should be done to prove the role of translational polymorphism in generation of new functional forms of eukaryotic proteins.

Acknowledgement

This work was supported by the Russian Foundation for Basic Research (02-04-48508) and RIKEN grant for bilateral research. We thank RAS programs (Dynamics of Plant, Animal and Human Gene Pools; Origination and Evolution of Biosphere; Molecular and Cellular Biology); and SB RAS Complex Integration Program (No. 59), Ministry of Education (PD02-1.4-464) and Ministry of Industry, Sciences and Technologies of Russian Federation (grants 43.106.11.0011 and Sc.Sh.-2275.2003.4) for partial support.

References

Emanuelsson O., Nielsen H., Brunak S., von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence // J. Mol. Biol. 2000. V. 300. P. 1005–1016.

Kettman J.R., Coleclough C., Frey J.R., Lefkovits I. Clonal proteomics: one gene – family of proteins // Proteomics. 2002. V. 2. P. 624–631.

Kochetov A.V., Sirnik O.A., Rogozin I.B., Glazko G.V., Komarova M.L., Shumny V.K. Context organization of mRNA 5'-untranslated regions of higher plants // Russ. J. Mol. Biol. 2002. V. 36. P. 510–516.

Kozak M. Pushing the limits of the scanning mechanism for initiation of translation // Gene. 2002. V. 299. P. 1–34.

Kumar A., Agarwal S., Heyman J.A., Matson S., Heidtman M., Piccirillo S., Umansky L., Drawid A., Jansen R., Liu Y., Cheung K.-H., Miller P., Gerstein M., Roeder G.S., Snyder M. Subcellular localization of yeast proteome // Genes Devel. 2002. V. 16. P. 707–719.

Leister D. Chloroplast research in the genomic age // Trends Genet., 2003. V.19, P.47–57.

Lukaszewicz M., Feuermann M., Jerouville B., Stas A., Boutry M. In vivo evaluation of the context sequence of the translation initiation codon in plants // Plant Sci., 2000. V.154, P.89–98.

Peltier J.-B., Emanuelsson O., Kalume D.E., Ytterberg J., Friso G., Rudella A., Liberles D.A., Soderberg L., Roepstorff P., von Heijne G., van Wijk K.J. Central function of the luminal and peripheral thylakoid proteome of Arabidopsis determined by experementation and genome-wide prediction // Plant Cell. 2002. V. 14. P. 211–236.

Rogozin I.B., Kochetov A.V., Kondrashov F.A., Koonin E.V., Milanezi L. Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a "weak" context of the start codon // Bioinformatics. 2001. V. 17. P. 890–900.

Watanabe N., Che F.-S., Iwano M., Takayama S., Yoshida S., Isogai A. Dual targeting of spinach protoporphyrinogen oxidase II to mitochondria and chloroplasts by alternative use of two in-frame initiation codons // J. Biol. Chem. 2001. V. 276. P. 20474–20481.

ANALYSIS OF PLANT MITOCHONDRIAL GENOME ORGANIZATION: CHARACTERISTICS OF REPEATS AND SEQUENCE COMPLEXITY

***Konstantinov Yu.M.*[1], *Poplavsky A.S.*[2], *Orlov Yu.L.*[2]**

[1] Siberian Institute of Plant Physiology and Biochemistry SB RAS, Irkutsk, Russia; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

***** Corresponding author: e-mail: yukon@sifibr.irk.ru; orlov@bionet.nsc.ru

Keywords: *mitochondrial genomes, plant genomes, low complexity regions, computer analysis*

Summary

Motivation: Plant mitochondrial genome analysis is one of pivotal tasks of modern structural analysis of complete genomes. It is of a special interest because higher plant mitochondria genomes have an enormous size (300–800 kb) as compared with the DNA of animal or fungal mitochondria (16–85 kb) and they are tolerant to incorporation of foreign DNA sequences of nuclear or chloroplast origin.

Results: Several numerical measures of text complexity including combinatorial, linguistic and Lempel-Ziv estimates were implemented in a software tool "LowComplexity" and "LZcomposer". We use the software developed for mitochondrial genomes analysis. The analysis of mitochondrial genomes in a few plant species showed that mitochondrial genomes have an average a less complexity sequences in comparison with nuclear chromosomes in the same species. The low complexity regions don't contain as a rule coding parts of genes. The low complexity regions in a gene coding parts contain RNA editing sites.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/.

Introduction

The structural and functional study of mitochondrial genomes in eukaryotes including plants and animals is of great importance. Mitochondria are well known as the cellular power factories. Much less is known about the structural and functional organization of genetic system in these organelles (Hoffmann *et al*., 2001). In plant mitochondria the pathway from the genetic information encoded in the DNA to the functional protein leads through a very diverse RNA world. How the RNA is generated and what kinds of regulation and control mechanisms are operative in transcription are current topics in research. Furthermore, the modes of posttranscriptional alterations and their consequences for RNA stability and thus for gene expression in plant mitochondria are currently objects of intensive investigations (Binder, Brennicke, 2003).

Therefore, the main goal of this work was the analysis of the whole mitochondrial (mt) genome sequences in plants to reveal some features in their structural organization potentially related to the functioning and evolution of plant mitochondria. These novel considerations may yield important clues for the further analysis of the plant mitochondrial genetic system.

Analysis of genomic sequences issues the challenge to search for the regions with the low text complexity, which could be functionally important (Hancock, 2002; Stern *et al*., 2001; Wan *et al*., 2003). Low complexity regions are often treated as the regions of biased composition containing simple sequence repeats (Hancock, 2002). As the method for complexity evaluation, we have chosen the scheme of the text representation in terms of repeats. The Lempel–Ziv may be interpreted as representation of a text in terms of repeats (Gusev *et al*., 1993; 1999). Based on this approach, we represent here the Internet-available tools LZcomposer (Orlov *et al*., 2004; http://

wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/) and LowComplexity (http://wwwmgs.bionet.nsc.ru/programs/low_complexity/).

We analyzed distribution of perfect repeats in plant mitochondrial genomes by the software developed. The mitochondrial genomes and chromosomes were studied in the following plant species: *Arabidopsis thaliana*, *Beta vulgaris, Oryza sativa, Brassica napus, Marchantia polymorpha.*

Method

Mitochondrial genomic sequences were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/). Complexity profiles for genome sequences were constructed by LowComplexity software. Perfect repeats were found by LZcomposer program.

The scheme of symbol sequence presentation by Lempel and Ziv was used to measure complexity of sequence by the number of steps of generating process (Gusev *et al*., 1999). The permitted operations here are generation of a new symbol (this operation is necessary at least to synthesize the alphabet symbols) and direct copying of a fragment from the already generated part of the text. Copying implies search for a prototype (repeat in a common sense) in the text and extension of the text by attaching the 'prepared' block. The algorithm implementation for DNA research was described in details in (Gusev *et al*., 1999).

The scheme for generating the sequence S may be represented as a concatenation *H* of the fragments:

$$H(S) = S[1 : i_1], S[i_1 + 1 : i_2], \ldots S[i_{k-1} + 1 : i_k], \ldots S[i_{m-1} + 1 : N], \quad (1)$$

where $S[i_{k-1} + 1 : i_k]$ is the fragment (component) generated at the k-th step (a sequence of elements located from the position $i_{k-1}+1$ to i_k); *N*, the length of sequence; and $m = m_H(S)$, the number of steps generating the process. The scheme with minimal number of steps m should be selected. This scheme determines the complexity of sequence S:

$$CLZ(S) = \min_H \{m_H(S)\}. \quad (2)$$

The minimal number of components in (1) is provided by selection at each step of the maximally long prototype in the previous history. The complexity decomposition of a sequence is performed from left to the right. The algorithm implementation for DNA research was described in details in (Gusev *et al*., 1999; Orlov *et al*., 2004).

One and the same sequence fragment may be composed of various types of repeats overlapping one another. Choosing of the longest prototype allows the maximally long perfect repeats to be found and specifies the standard of their representation as complexity decomposition. A small improvement of the program allows long imperfect repeats to be detected in the text. The result of program running is sequence of the components $S[i_{k-1} + 1 : i_k]$, their lengths, positions of the prototypes, and the way of copying (i.e., type of repeat).

In Figure 1, there is an example of complexity decomposition of the nucleotide sequence containing the G->U RNA editing sites, AGAGAGTCCCACATACGAGA. The components of complexity decomposition are separated by dots. Black and gray arrows below the sequence mark the copied fragments and their prototypes. Tandem repeat characterized by partial overlapping of the prototype on the copied fragment is marked by dotted line. In decomposition given above, the first one-lettered components, A and G, are composed by operation generating novel symbol. The complexity of this 20-lettered sequence equals to 11 (the number of components in H(S)).

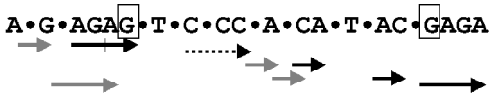


Fig. 1. Example of complexity decomposition by modified Lempel-Ziv method. Repeats are separated by bold dots. Black and gray arrows designate repeated fragments and their prototypes correspondingly. RNA editing positions (G->U) marked by rectangles.

If there are some alternative variants of copying, the program applies the prototype, which is the nearest to the component synthesized.

We construct the complexity profile in the sliding window with the length N, the evaluation of complexity is calculated as the whole number *CLZ*(S) of components of complexity decomposition in the window N, or as the relative number of the components *CLZ*(S)/N (For the example in Figure 1 normalized complexity value *CLZ*(S)/N=11/20=0.55).

The program for complexity evaluation in accordance with Lempel-Ziv method with extended attenuation is available at http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/.

We use also linguistic complexity estimation (Troyanskaya *et al*., 2002) and entropy profile (Wootton, Federhen, 1996) to search low complexity regions in the genomes.

Results and Discussion

Let us consider complexity decomposition of mitochondrial genomes. We present a sequence as a concatenation of non-overlapped fragments, every such fragment has a prototype (direct or inverted copy) in upstream sequence, and every such fragment has maximal possible length. We study distribution of repeated fragments and their types (direct (*D*) or inverted(*I*)).

The Table shows that significant part of genome is populated by long exact repeats. Direct long exact repeats with length greater than 20 bp are more preferable. Longest perfect repeats belong often to inverted ones.

Figure 2 shows the complexity profile of mitochondrial genome in *Arabidopsis thaliana.*

Table.

Mt genome	Size, Kbp	Part of genome size covered by long (>20bp) exact repeat			Mean repeat length in decomposition	Type of maximal repeat	Maximal repeat size, bp
		All	Direct	Inverted			
A. thaliana	351	0.016	0.009	0.007	9.17	*I*	6,590
B. vulgaris	332	0.014	0.008	0.006	9.15	*I*	19,646
O. sativa	283	0.007	0.004	0.003	8.92	*D*	43,760
B. napus	217	0.016	0.008	0.008	8.80	*D*	1,758
M. polymorpha	184	0.024	0.014	0.010	8.84	*I*	428

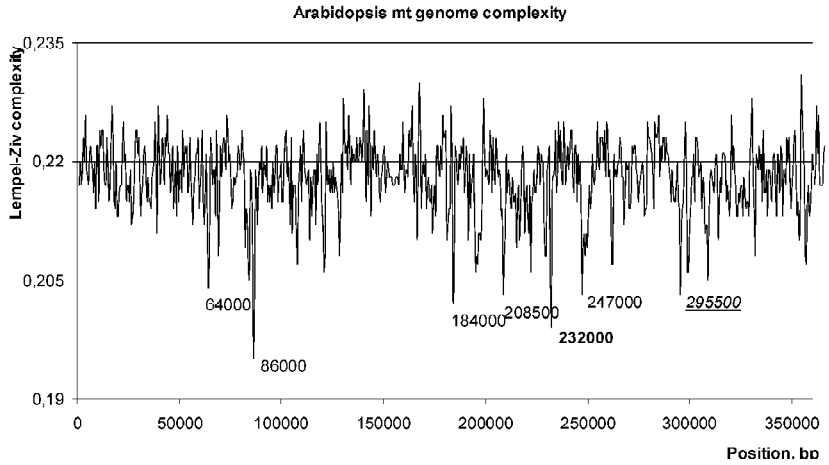


Fig. 2. Complexity profile for *A.thaliana* mt genome in sliding window 1000 bp. Profile uses Lempel-Ziv (LZ) complexity estimation. Positions of low complexity regions are indicated by labels.

The complexity was calculated using scale [0;1]. Complexity profile in sliding window 1000 bp was calculated by LowComplexity program. Minimal values of complexity profile are indicated (less than 3*σ). These values correspond to low complexity regions. The minimal values correspond regions that do not contain genes. Region [231,500; 231,500] containing cytochrome C biogenesis orf382 is marked by bold font. This gene contains several RNA editing sites.
Low complexity regions in *Beta vulgaris* (sugar beet) mt genome contain tRNA genes. Other mt genomes also shown similar distribution of large low complexity regions.

Our analysis of genomic DNA shown high correlation of Lemple-Ziv and linguistic measures of text complexity. Entropy measures have less correlation with Lempel-Ziv estimation for large sliding window size (>1Kb). At average mt genomes have less complexity by Lempel-Ziv measure than such complexity estimations for chromosomal sequences. At the same time entropy of mt genome is higher. This observation suggests greater presence of long repeats in mitochondrial genomes.

Plant mitochondrial genomes have complex structure. In mitochondrial genome of *Arabidopsis* sequences of nuclear origin represent 4% of the sequence. About 2 % of the genome is composed of unaccounted sequences found both in the nucleus and the mitochondria and 1.2 % of the genomic DNA are sequences of plastid origin.

Let us note great size of maximal exact repeat 43.7Kb in rice mt genome. This value is unique for sequences of such size. (Only two bacterial genomes have exact repeats greater than 40Kb – *E. coli* (strain O157 H7) and *Streptococcus agalactiae*, strain NEM316). See complete table at http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/ResBacterial.htm.

Acknowledgements

This work was supported in part by the RFBR (01-07-90376, 02-07-90355, 03-04-48506), Russian Ministry of Education (E 02-6.0-250), NATO (LST.CLG 979815) and SB RAS (Integration project No. 119).

References

Binder S., Brennicke A. Gene expression in plant mitochondria: transcriptional and post-transcriptional control // Philos. Trans. R Soc. Lond. B Biol. Sci. 2003. V. 358(1429). P. 181–8.

Hoffmann M., Kuhn J., Daschner K., Binder S. The RNA world of plant mitochondria // Prog. Nucleic Acid Res. Mol. Biol. 2001. V. 70. P. 119–54.

Gusev V.D., Kulichkov V.A., Chupakhina O.M. The Lempel-Ziv complexity and local structure analysis of genomes // Biosystems. 1993. V. 30. P. 183–200.

Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences // Bioinformatics. 1999. V. 15. P. 994–9.

Hancock J.M. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects // Genetica. 2002. V. 115(1). P. 93–103.

Marienfeld J., Unseld M., Brennicke A. The mitochondrial genome of Arabidopsis is composed of both native and immigrant information // Trends Plant Sci. 1999. V. 4(12). P. 495–502.

Orlov Yu.L., Gusev V.D., Miroshnichenko L.A. LZcomposer: decomposition of genomic sequences by repeat fragments // Biophysics (Mosk.), 2004, (in Press).

Orlov Yu.L., Gusev V.D., Nemytikova L.A. Software package LZcomposer: analysis of occurrence of repeats in complete genomes // Proc. of BGRS'2002. Novosibirsk: Inst. of Cytology&Genetics Press, 2002. V. 3. P. 247–250.

Stern L., Allison L., Coppel R.L., Dix T.I. Discovering patterns in Plasmodium falciparum genomic DNA // Mol. Biochem. Parasitol. 2001. V. 118(2). P. 175–86.

Troyanskaya O.G., Arbell O., Koren Y., Landau G.M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity // Bioinformatics. 2002. V. 18(5). P. 679–88.

Wan H., Li L., Federhen S., Wootton J.C. Discovering simple regions in biological sequences associated with scoring schemes // J. Comput. Biol. 2003. V. 10(2). P. 171–85.

Wootton J.C., Federhen S. Analysis of compositionally biased regions in sequence databases // Methods Enzymol. 1996. V. 266. P. 554–71.

MANUAL CURATION OF EST LIBRARIES BY TISSUE SPECIFICITY AND CELL ORIGIN

***Kosmodemiansky Il.A.*[1]*, *Gelfand M.S.*[1,2,3], *Mironov A.A.* [1,3]**

[1] Department of Bioengineering and Bioinformatics, Moscow State University; [2] Institute for Information Transmission Problems, RAS, Moscow; [3] State Scientific Center GosNIIGenetika, Moscow
* Corresponding author: e-mail: ilya@integratedgenomics.ru

Keywords: *EST, alternative splicing, tissue specificity, tumor expression*

Summary

Motivation: Accurate manual curation of EST libraries from NCBI/UniGene by tissue specificity and disease state is an important aspect of modern computational molecular biology. Such resource would provide a powerful tool for investigators working in the area of tumor biology and alternative splicing.

Results: We curate about 2500 EST libraries by disease state and cell origin (cell line or a donor).

Availability: www.ig-msk.ru:8005/EDAS/

Introduction

Alternative splicing has recently emerged as a major mechanism of generating the protein diversity in higher eukaryotes. Identification of proteins (or mRNAs) specifically expressed in tumors is a major challenge of tumor biology and molecular medicine. A particular case of such analysis is searching for tissue-specific and tumor-specific patterns of alternative splicing.

Expressed Sequence Tags (EST) provide a powerful resource for analysis of alternative splicing (Nurtdinov *et al*., 2003). Thus the information about tissue from which a EST library has been derived can be used to determine tissue specificity of splicing isoforms. The problem is, that UniGene is not curated database (Wheeler *et al*., 2001) and we have no criteria to believe posted information about tissue library derived from. So we need to perform such manual curation (let call it classification).

Classification of EST libraries from the NCBI/UniGene database by tissues and disease state (cancer or normal) was done in LibRegistry by Baranova *et al*. (2001). However, the number of EST libraries in UniGene has increased considerably since 2001 (currently it is approximately 7000). Further, in that study the samples from cel lines and from primary cancer tissue were not separated.

In the EDAS project (EST-Derived Alternative Splicing Database) (Nurtdinov, Kosmodemiansky, 2003) we created a catalogue of EST libraries. This catalogue contains information about clone libraries, tissues and organs from which they have been derived, and their donor or cell line origin.

Methods

Primary information about clone libraries was obtained from UniGene (www.ncbi.nlm.nih.gov/UniGene) using perl script-based query (all perl scripts are available upon request). Then this information was parsed also using perl scripts. Libraries were identified as normal ones based on explicit information from UniGene. If this information was missing or contradictory, manual analysis was performed using information from the CGAP (Cancer Genome Anatomy Project of the National Institutes of Health) web site (http://www.ncbi.nlm.nih.gov/CGAP/). We also searched PubMed for articles where the library construction procedures were described. Finally, we used

Google search to find information about commercial cell lines at web sites of biotechnological companies.

Then we classified the clone libraries into the cell line/donor categories. The same procedure was applied: explicit descriptions were derived and used "as is", whereas manual parsing was used to determine the disease state using indirect indication in other description fields (for example, age=65 likely means "donor").

Results

The most recent version of the EST-library catalogue is available upon request (as a MS Excel or a GNU OpenOffice/calc file) or as a part of the EDAS database (www.ig-msk.ru:8005/EDAS/). Correctly it contains the data about 2500 libraries, including all libraries of the germ line origin (testis, ovary) and also placenta. By the end of 2004, we plan to annotate 90 % of UniGene libraries covering (about 90 %) of all ESTs.

Acknowledgements

We are grateful to Ramil Nurtdinov for useful discussion. This study was partially supported by grants from the Howard Hughes Medical Institute (55000309), the Ludwig Institute for Cancer Research (CRDF RB0-1268), and the Russian Found for Basic Research (04-04-49440).

References

Baranova A.V. *et al*. In silico screening for tumor-specific expressed sequences in human genome // FEBS Lett. 2001. V. 508(1). P. 143–8.

Nurtdinov R.N., Artamonova I.I., Mironov A.A., Gelfand M.S. Low conservation of alternative splicing patterns in the human and mouse genomes // Hum. Mol. Genet. 2003. V. 12(11). P. 1313–20.

Nurtdinov R.N., Kosmodemiansky I.A. EDAS – EST derived alternative splicing database // Proc. of MCCMB'03, Moscow, 2003.

Wheeler D.L. *et al*. Database resources of the National Center for Biotechnological Information // Nucl. Acids Res. 2001. V. 29. P. 11–16.

DETECTION OF CLASSICAL ATTENUATION IN BACTERIAL GENOMES

Leontiev L.A.*, Shirshin M.A., Lyubetsky V.A.

Institute for Information Transmission Problems, RAS, Moscow, Russia
* Corresponding author: e-mail: leontiev@iitp.ru.

Keywords: *attenuation regulation, threonyl-tRNA synthetase, tryptophan biosynthesis, beta-proteobacteria, branched amino acid biosynthesis, Thermus/Deinococcus group*

Summary

Motivation: Screening of the genomes for signals for patterns of gene expression regulation including attenuation with the leader peptide, is an important task in itself and in the general context of bioinformatics, for instance, in the field of gene annotation. However, development of effective algorithms for mass attenuation detection is is far from complete(for details see, in particular, Lyubetskaya, Molecular Biology 2003; Vitreshak *et al*., 2004). The abstract presents the second part of a collaborative work first described in the conference proceedings in (Lyubetsky *et al*., 2004).

Results: Novel cases of attenuation regulation of the aminoacyl tRNA synthetase, branched amino acid and tryptophan biosynthesis are detected in the beta-proteobacteria, Actinobacteria and Thermus/Deinococcus groups.

Introduction

The "classical" attenuation prototype includes the leader peptide reading frame with a number of regulatory codons and the three hairpins – pause hairpin, antiterminator and terminator with a run of Us (Lyubetskaya *et al*., 2003; Vitreshak *et al*., 2004, in press). We found that such aregulation usually requires three conserved fragments (which we call "three words"), like the bases of the hairpins. Pairing of the first and second word switches on the antiterminator, pairing of the second and third word – the terminator. Supposedly, alternative blocking of the first word by the ribosome determines which one of the alternative conformations is formed.

Methods and Algorithms

Nucleotide sequence data were obtained from the NCBI databases. Putative attenuation structures were detected with an ad hoc algorithm described in (Gorbunov *et al*., 2001; Lyubetskaya *et al*., 2003).

Implementation and Results

1. We predicted attenuation regulation of genes involved in branched aminoacid biosynthesis in the Thermus/Deinococcus group of gram-positive bacteria. In *Deinococcus radiodurans*, attenuation of the gene putatively involved in leucine biosynthesis was found. The regulated gene (having the attenuation structure within the leader segment) named *leuA2* after a homologous gene in *Sinorhizobium meliloti* (homologs are also found in alpha-proteobacteria, Actinobacteria and fungi) encodes the 2-isopropylmalate synthase and is located at the beginning of a 4-gene operon. The other three genes do not have homologs in alpha- and gamma-proteobacteria and *Bacillus subtilis*. In *Thermus thermophilus,* the closest relative of *Deinococcus radiodurans*, three 2-isopropylmalate to synthase isozymes were found but none contains an open reading frame with a set of leucine codons. The attenuation structure is shown below.

Deinococcus radiodurans leuA2. Start codon is atposition 1496949 in the genome. Start, stop and regulatory codons are set in bold, the antiterminator underlined, terminator set in the uppercase.

atgcctgcatacgg**cctgcttcttctt**agc**ctt**cctccggtggag**tga**cggcggacgcagcccgtatcccagCCCCCGGAGGa
ttcagaCCTTCCGGGGGacatttttgt

In addition, the attenuation-regulated operon *ilvBN-x-ilvC* was found in *Deinococcus radiodurans*, where the *ilvBN* genes encode two subunits of acetolactate synthetase, the function of the gene *x* is unknown (its length being 368 bp), and *ilvC* encodes ketol acid reductoisomerase.

Deinococcus radiodurans ilvBN-x-ilvC. Start codon is at position 1531471 in the genome.

atgagcgctggaaac**gtgattgtacttctt**agccacgggggacaccgaggc**taa**gcaggttgtaccggtttcagtcgcACCCCCC
GCCCaaaccaaGGAGGCGGGGGGTttt

Three genes, *ilvGM, ilvBN, ilvIH*, encoding the acetolactate synthase isozymes were found in Enterobacteria, with the first two attenuation regulated, while only one acetolactate synthase was found in *Deinococcus radiodurans*. The same situation is with *Thermus thermophilus* (the genome is not publicly available), which contains only one acetolactate synthase IlvBN. The synthase is located within the *ilvBNC-leuA2-x-leuA2* operon. The genes encode: two acetolactate synthase subunits, ketol-acid reductoisomerase, 2-isopropylmalate synthase LeuA2, ribosome alanine-transferase and 2-isopropylmalate synthase LeuA2' again; *leuA2* and *leuA2'* are 30 % homologous.

2. This article also describes novel attenuators located upstream of the genes of tryptophan and tryptophanyl-tRNA synthetase biosynthesis in some Actinobacteria. *Streptomyces avermitilis* and *Corynebacterium diphtheriae* possess genes *trpE1* and *trpE2* that are homologous to *trpE* of *E. coli*. In *Streptomyces avermitilis*, the *trpE2* gene is a part of an operon uniting genes with unknown function, its homolog in *Corynebacterium diphtheriae* is located in the *x-trpB-trpE2-trpGDC* operon*;* expression of both genes is not attenuation regulated. However, the *trpE1* genes encoding an anthranilate synthase isozyme are separated within the genomes of both organisms, regulated by attenuation.

Streptomyces avermitilis trpE1. Start codon has position 7322414 in the genome.
atgttcgcgcactcgatccagaac**tggtggtgg**accgctcatccggcggcccac**tga**ctgcgcgtacgcaagacttcgcgAAGGCCGCC
CgagGGGCGGCCTTtcgtgtt
Corynebacterium diphtheriae trpE1. Start codon has position 2456514 in the genome.
atgaatgcacataac**tggtggtgg**cgcgct**taa**ccgcggggccgttttcacgcattcatttcaacAGGCTCgc
CTTGTccaACAAGCAGCGGGCCTttttgtta
Also in *Streptomyces avermitilis*, the *trpS2* gene encoding the tryptophan-tRNA synthetase is probably regulated by attenuation each.
Streptomyces avermitilis trpS. Start codon is at 5758647 in the genome.
atgactacgcgtacgtgtacccagcag**tggtgg**gccgcc**tga**cggcggccgtacacacgtatgtactcAACGGCCGCCG
cctCGGCGGCCGTTctcgttt

This organism contains two tryptophan-tRNA synthase isozymes encoded by *trpS* and *trpS2,* both genes are homologous to *trpS* of *Escherichia coli* and *Corynebacterium duphtheriae*. Notably, the *trpS* genes in *Escherichia coli* and *Corynebacterium diphtheriae* do not have paralogs and are probably not regulated by attenuation.

3. Putative attenuators of threonyl-tRNA synthetase were predicted in beta-proteobacteria, namely, in *Bordetella* spp. (*Bordetella bronchiseptica*, *Bordetella parapertussis*, *Bordetella pertussis*), *Ralstonia* spp. (*Ralstonia metallidurans*, *Ralstonia solanacearum*), *Chromobacterium Vilaceum,* and in *Methylococcus capsulatus,* which is tentatively classified within gamma-proteobacteria. In *Bacillus subtilis*, two isozymes of the threonyl tRNA synthase were found, *thrS* and *thrZ*, which are 59 % homologous. The first, *thrS*, is not regulated by attenuation, and the second one, *thrZ*, contains a threonine-dependent attenuator.

Bacillus subtilis thrZ. Start codon is at 3855364 in the genome.

atgctgcgttacagcaccgagccgacaaca**tag**tatttgtcgggaacttgggtggaaccacgggttaatcACACACTCGTCCC
TATCTGCGGGACGGGTGTGTtttttta

In *Escherichia coli*, only one *thrS* gene was found. The gene encodes threonyl-tRNA synthase

and is not attenuation regulated. Attenuation structures in *Bordetella* spp. are completely identical in sequence. The one from *Bordetella pertussis* is shown below.

Bordetella pertussis thrZ. Start is at1574598 in the genome.

atgctgctgcgcccgacacgaactacgaccggaatcttccgcactca
gtcgcgt**taa**tgatttcatgcgtcgggttggtatcgcgtcattgcggtaaatactcaacaaggcaccaGACAAAACGCGGC
caggcaGCCGCGTTTTTCgtttcc

Attenuation structures in *Ralstonia solanacearum*, *Chromobacterium vilaceum* and *Methylococcus capsulatus* are also shown below.

Ralstonia metallidurans Reut_370 (starts at 108811)

atgagcaagacaactcgaactactaccgctggt**tag**cgacagtagtcga
ggtgcgccttcgacccgtagtgtgtgatacggggaaacacagaaaaACGCGGCC
ttGGCCGCGTttttttcgtc

Ralstonia solanacearum (starts at 1692246)

atgatccaggcaccgcgaactaccaccgcttcggagcgacag**tag**tcaaggt
gcggtttcgccctttgcggtctgcccaagcggcagttcagcgcaaatgaaaACGC
GGCCttGGCCGCGTttttttcgtccgcgt

Chromobacterium Vilaceum ATCC 12472 (starts at 1420122)

atggtgggggtcgttggttcgaatccaatcgtgcctaccaaattaaatgcatggagttgccaagtgaaatccttgcgaactaccacaacccgacatactcacgaaa
gc**tga**gccttgcgggttggaccatgcgaaaaAGAAGTGCGGCTcaGGCCGCACTTTTtttttaccctt

Methylococcus capsulatus str. Bath (starts at 729164)

atgcctccttcttgtcgaattacgacaggcgcg**tag**ctcagt
tggttagagcaccaccttgacatggtgggggtcgttggttcgagtccaatcgcgcctaccagatatcccagGAAGCATCGCCa
GGCGATGCTTTtttgtttggggg

Discussion

Novel cases of attenuation regulation of biosynthesis of several aminoacyl-tRNA synthetases, branched amino-acids and tryptophan are found in bacteria *Deinococcus radiodurans, Thermus Thermophilus, Streptomyces avermitilis, Corynebacterium diphtheriae* and the groups of beta-proteobacteria Bordetella spp, Ralstonia spp, *Chromobacterium vilaceum, Methylococcus capsulatus*. The regulation patterns are partly described above and dealt with in in more detail in the presentation.

The conventional attenuation is a universal mechanism of gene expression regulation in terms of both its wide occurrence across the organismal diversity (it is reported for many bacterial taxa from Bacteroidetes /Chlorobi, Firmicutes, Thermus/Deinococcus, Actinobacteria, *etc.*) and functional diversity of the regulated operons (amino acid, animoacyl-tRNA synthetases, transporters biosynthesis). Interestingly, the threonyl-tRNA synthetase regulation in *Methylococcus capsulatus* putatively classified within the gamma-proteobacteria is similar to the analogous mechanism in beta-proteobacteria, while within the other gamma-proteobacteria it is not reported.

Acknowledgments

We are grateful to M.S. Gelfand, D.A. Rodionov, A.G. Vitreshak and A.V. Seliverstov for helpful discussions, as well as to V.V. Zubov for providing us with some biological data.

References

Gorbunov K.Yu., Lyubetskaya E.V., Lyubetsky V.A. On two algorithms of detection of alternative elements of RNA secondary structure // Informational processes. 2001. V. 1(2). P. 178–187.

Lyubetskaya E.V., Leontiev L.A., Gelfand M.S., Lyubetsky V.A. Search for alternative secondary tRNA structures, regulating bacterial gene expression // Mol. Biol. 2003. V. 37, N 5. P. 834–842.

Lyubetskaya E.V., Leontiev L.A., Lyubetsky V.A. Alternative secondary structure detection in a group of gamma-bacteria // Informational processes. 2003. V. 3(1). P. 23–38.

Lyubetsky V.A., Seliverstov A.V. Amino Acid Biosynthesis Attenuation in Bacteria // This conference. 2004.

Vitreshak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis // FEMS Microbiology Letters. Accepted, 2004. 39 p.

ANALYSIS OF THE CONTEXT FEATURES OF SF-1 BINDING SITE AND DEVELOPMENT OF A CRITERION FOR SF-1 REGULATED GENE RECOGNITION BY THE SITEGA METHOD

Levitsky V.G.*, Ignatieva E.V., Busygina T.V., Merkulova T.I.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: levitsky@bionet.nsc.ru

Keywords: *steroidogenic genes search, SF-1 site recognition, discriminant analysis.*

Summary

Motivation: Development of methods allowing search for transcription factor binding sites (TFBSs) is important in investigation of the regulatory regions in the eukaryotic genes. The currently used methods are not accurate enough to recognize the binding sites of a transcription factor (TF) or a group of factors of a single family.

Results: We propose a new method to search for TFBSs exemplified by the SF-1 binding site. The approach involves partitioning the regions including TFBSs into local regions and choice of the most significant dinucleotides for each. The SF-1 site recognition method is used for developing a criterion for the steroidogenic genes search. The method was tested on samples of the gene regions from the TRRD database. We predicted 6 new genes for which direct or indirect evidence supports the idea that they are components of the steroidogenic system.

Introduction

Recognition of TFBSs by computer methods is an effective approach for the regulatory gene regions search. The SF-1 factor belongs to the nuclear receptor family, and is a key regulators of steroidogenic gene expression in gonads and adrenals (Val *et al*., 2003). The SF-1 is required for development and function at all levels of the hypothalamic-pituitary-gonadal and adrenals axis (Val *et al*., 2003). At present, there is experimental evidence for the presence of the SF-1 binding sites in the regulatory regions of many genes functioning within this axis (Busygina *et al*., 2003). Nevertheless, an overall pattern of SF-1 mediated regulation is far from complete.

Methods and Algorithms

The SF-1 sites with flanks (93 bp in length) with centrally located TFBS were extracted from the TRRD database (SITES sample, Table 1).

Table 1. Samples of nucleotide sequences

Sample name	Location with respect to transcription start	Number of sequences
SITES	from – 4830 to +140	54
EMBL(STER)	[-500; +1]	33
TRRD(STER+)	[-2100;+2100][1]	58
TRRD(STER-)	[-2100;+2100][1]	1274

[1] – lacking the 5'- or 3'-flanks of nucleotide sequences completed with the symbol 'n'.

A control sample of promoter regions of the genes controlling steroidogenesis EMBL(STER) was used for the choice of the threshold in building the recognition methods. The sample was composed of genes for which the fact of SF-1 regulation and the presence of SF-1 sites in regulatory regions was not investigated experimentally. The sample was set up directly from the EMBL database using the literature sources. To develop a criterion for the detection of the SF-1 regulated genes, we used two groups of genes described in the TRRD database (Table 1): (i) genes controlling steroidogenesis for which the presence of the SF-1 sites has been demonstrated experimentally (TRRD(STER+)); (ii) the remaining genes which lack the experimentally approved SF-1 sites (TRRD(STER-)).
The SiteGA method was implemented using of a genetic algorithm involving a discriminant function of a local dinucleotide context characteristics. Firstly, the method divides the entire analyzed region (93 bp) into three overlapping 36 bp long stretches (the left, central, right). At the second step, for each stretch the optimum partition was searched and the most significant dinucleotide characteristics were chosen, as described (Levitsky, Katokhin, 2003). At this step, the recognition function $\varphi_r(X_r,\alpha_r)$ for each stretch (r = 1,2,3) and significance level α_r was found. To recognize a site in a nucleotide sequence, a 93 bp long sliding window (X) and the corresponding division $\{X_1,X_2,X_3\}$ were used. For each window position, the recognition function value was calculated from the equation:

$$\varphi(X, \alpha_1, \alpha_2, \alpha_3) = \begin{cases} 0, \text{ if } \varphi_r(X_r,\alpha_r) < 0, \text{ for one of stretches, } 1 \le r \le 3 \\ (1/3) * \sum_{r=1}^{3} \varphi_r(X_r, \alpha_r) \end{cases} \quad (1)$$

The type I and type II errors for the SF-1 site was estimated by BootStrap method (Table 2). At the third step, the optimum significance levels for the recognition functions of 3 stretches were searched. These significance levels were found using two control samples of sequences: (i) EMBL(STER) (Table 1) and (ii) random sequences with the same nucleotide content as in the SITES sample. Based on the predicted site ratio in the two control sequence samples the significance levels 0.95, 0.8, 0.7 for the central, right and left stretches, respectively, were chosen for further analysis. Type I error calculated for this set for the SITES sample was 50 %, while type II error for the control sample of random sequences was 1.09E-05 (1/91558).

Table 2. Type 1 and type II errors for SF-1 recognition by BootStrap

Type I error	0.51	0.74	0.8
Type II error	4.31E-03 (1/232)	3.61E-04 (1/2771)	9.09E-05 (1/10999)

Implementation and Results

Fig. 1 presents diagrams of significant correlations used for construction SiteGA method of SF-1 site recognition for the left flank, center and right flank. For example, GG dinucleotide frequency in the [25; 33] region positively correlated with the CT dinucleotide frequency in the [4; 14] region and negatively with the GC dinucleotide frequency in the [15; 24] region.

The distribution of the predicted site density for the TRRD gene regions, belonging to steroidogenic system (TRRD(STER+)) and the remaining genes for which this function has not been demonstrated (TRRD(STER-)) is shown in Fig. 2. It is noticeable that the main difference between the steroid regulated promoters and the non-steroid regulated is a sharp increase in site density in the [-300; +1] region. The density is 1.91/1000 bp for the (TRRD(STER+) sample and 0.11/1000 bp for the (TRRD(STER-) sample.

Based on the distribution of the density of the predicted SF-1 sites for the TRRD gene regions (Fig. 2), we suggest the following criteria for predicting genes regulating steroidogenesis: 1) the presence of at least one SF-1 site in the [-300; +1] region; 2) the presence of at least one SF-1 site in the [-2100; -300] or [+1; +2100] regions if criterion 1 is met. As expected (Table 3), in the TRRD(STER+) sample containing SF-1 regulated genes, the percentage for the genes meeting the suggested criterion 2 (15.5 %), is considerably greater than that for the TRRD(STER-) sample (1.1 %). The criterion 2 as more severe was used for predicting new genes regulating steroidogenesis.

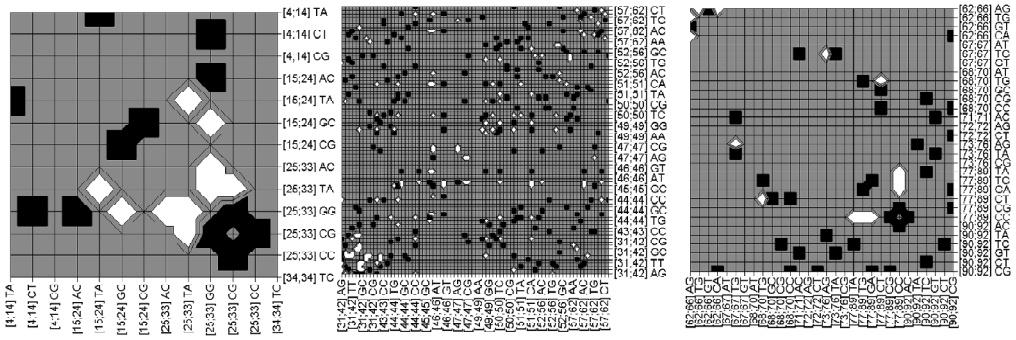


Fig. 1. Significant correlations between dinucleotide frequencies ($p < 0.05$), used for the construction the SiteGA method of SF-1 site recognition for left flank (A), center (B) and right flank (C). The significant positive correlations are black and the significant negative are white, grey indicates no correlation.

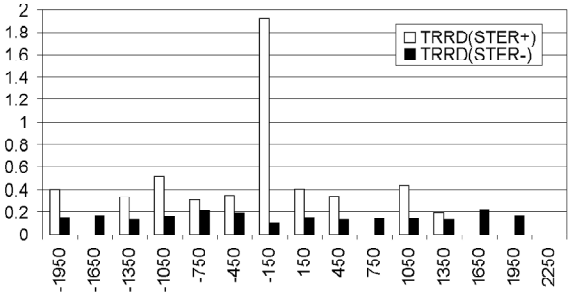


Fig. 2. The density of the predicted SF-1 sites for gene regions, belonging to steroidogenic system (TRRD(STER+)) and the remaining genes (TRRD(STER-)). X axis – the position of the interval center relative to the transcription start site, Y axis – the number of predicted sites per 1000 bp.

Discussion

The analysis of the significant correlations between dinucleotide frequencies in the local regions of the SF-1 site (Fig. 1) demonstrated that the majority of correlations is positive (132, 62 %); the great majority of correlations is in the center (174, 82 %); the right flank was found to be more informative than left (27 correlations against 11). The occurrences of significant correlations for SF-1 site recognition out of the central 36 bp stretches was very considerable. The patterns beyond canonical footprint (no longer than 20 bp) reflect the genome nucleotide context around the site whose consideration helps to increase recognition accuracy.

The SF-1 recognition presented here was confirmed experimentally and it appears to be very accurate (Ignatieva *et al.,* 2004). The newly identified genes (2nd criterion, Table 3) are promising

Table 3. Application of criteria to the TRRD genes samples

Sample	Criterion 1	criterion 2
TRRD(STER+)	25 (43.1%)	9 (15.5%)
TRRD(STER-)	38 (2.9%)	15 (1.1%)

candidates for experimental checking. The product of the rat NO inducible nitric oxide synthase (iNOS) gene plays an important role at all the levels of the hypothalamic-pituitary axis. Furthermore there is experimental evidence for the involvement of the SF-1 factor in mouse iNOS transcription regulation (Wei *et al.,* 2002). There are 2 genes with similar function identified both in human and mouse. These genes encode transport proteins, the cellular retinol-binding protein II and retinol binding protein 3, interstitial, whose relation to steroidogenic processes has been demonstrated (Clarke, Armstrong, 1989). Also, a very important candidates are the rat D-2 dopamine receptor gene and human small proline-rich protein 1A. Experimental checking is of the great interest because the dopamine receptor regulates steroidogenesis in the adrenals (Morra *et al.,* 1992) and SF-1 expression in keratinocytes has been demonstrated (Val *et al.,* 2003).

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grants Nos. 02-04-48802, 03-04-48555, 03-04-48829, 01-07-90376, 02-07-90355, 03-07-96833, 03-07-90181); Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); SB RAS (integration project No. 119); NATO (grant LST.CLG.979816), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”.

References

Busygina T.V., Ignatieva E.V., Osadchuk A.V. Consensus sequence of transcription factor SF-1 binding site and putative binding site in the 5' flanking regions of genes encoding mouse steroidogenic enzymes 3betaHSDI and Cyp17 // Biochemistry (Mosc). 2003. V. 68(4). P. 377–384.

Clarke S.D., Armstrong M.K. Cellular lipid binding proteins: expression, function, and nutritional regulation // FASEB J. 1989. V. 3(13). P. 2480–2487.

Ignatieva E. V., Oshchepkov D.Yu., Levitsky V.G., Vasiluev G.V., Klimova N.V., Busygina T.V., Merkulova T.I. Comparison of the results of search for the SF-1 binding sites in the promoter regions of the steroidogenic genes using the SiteGA and SITECON methods // 2004. (this issue).

Levitsky V.G., Katokhin A.V. Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis // In Silico Biol. 2003. V. 3. P. 81–87.

Morra M., Leboulenger F., Vaudry H. Characterization of dopamine receptors associated with steroid secretion in frog adrenocortical cells // J. Mol. Endocrinol. 1992. V. 8(1). P. 43–52.

Val P., Lefrancois-Martinez A.M., Veyssiere G., Martinez A. SF-1 a key player in the development and differentiation of steroidogenic tissues // Nuclear Rec. 2003. V. 1. P. 8–45.

Wei X., Sasaki M., Huang H., Dawson V.L., Dawson T.M. The orphan nuclear receptor, steroidogenic factor 1, regulates neuronal nitric oxide synthase gene expression in pituitary gonadotropes // Mol. Endocrinol. 2002. V. 16(12). P. 2828–2839.

DNA NUCLEOSOME ORGANIZATION OF THE FUNCTIONAL GENES REGIONS AND ITS RELATION TO GENE EXPRESSION LEVEL

***Levitsky V.G.*[1], *Pichueva A.G.*[1], *Kochetov A.V.*[1], *Milanesi L.*[2]**

[1] Institute of Cytology & Genetics SB RAS, Lavrentieva 10, Novosibirsk, 630090, Russia; [2] Istituto di Tecnologie Biomediche Avanzate, via F. Cervi 93, 20090 Segrate, Milano, Italy
* Corresponding author: e-mail: levitsky@bionet.nsc.ru

Keywords: *gene expression pattern; nucleosome positioning*

Summary

Motivation: DNA nucleosome organization is an important factor in gene expression patterning. The nature of the context signals determining nucleosome formation sites are not completely understood. Given these considerations, the relation between nucleosome positioning and gene expression regulation appears of great interest.

Results: Taxon-specific nucleosome organization of the yeast and mammalian core promoters was identified. Positive correlations were established between characteristics potentially related to nucleosome positioning of the yeast core promoter and gene expression level. The context parameters of the DNA nucleosome organization differ by the distribution pattern in the mammalian and yeast promoters.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/, http://wwwmgs.bionet.nsc.ru/mgs/programs/phase/.

Introduction

DNA packaging in chromatin structure is an important factor in the regulation of the status of gene expression pattern in eukaryotes. The nucleosome positioning in certain circumstances may both activate and repress gene expression. It is now known that nucleosome formation in the genomic DNA is determined by the interaction of different regulatory and structural proteins with their cognate sites and context nucleosome positioning signals. Taken together, the interactions form the code of chromatin nucleosome organization (Lowary, Widom, 1997; Kiyama, Trifonov, 2002).

It may be assumed that nucleosome formation depends on the gene expression level and highly expressed genes have specific features, which increase transcription efficiency by optimization of context characteristics. Our previous observations demonstrated significant correlations between DNA context of the promoter and 5'UTR regions and the expression level (Kochetov *et al.,* 2002; Pichueva *et al.,* 2004). We have also previously shown that the nucleosome formation potential is greater in tissue-specific gene promoters than in the genes expressed in many tissues and housekeeping genes (Levitsky *et al.,* 2001). Hence, the capability of nucleosome positioning in the promoter region may serve as a regulatory factor of the gene expression level. Here, we present the results of a computational analysis of the characteristics potentially related to nucleosome positioning in the core promoters of the yeast and mammalian genes.

Materials and Methods

Sacchcaromyces cerevisiae mRNAs were extracted from the EMBL nucleotide sequence databank. Full-size 5'UTRs were selected from the entries containing a description of the transcription start sites (TSS) and complete coding regions. Only the mRNAs with full size 5'UTRs (i.e., containing reference to an experimentally mapped TSS) were used. This resulted in a set of 5'UTRs of 240 yeast genes. To avoid bias due to redundant sequence data in statistical analysis, redundant sequences (CDS homology higher than 70 %) were removed. Finally, the set comprised 5'UTRs of 98 yeast genes with a single TSS and a complete coding sequence. A sample of promoter sequences spanning 150 nucleotides upstream of the major TSS was also compiled and was

combined with the corresponding 5'UTRs; 271 promoter sequences of length 1500 bp ([–1000; +500] with respect to the TSS) of the eykaryotic genes (mostly mammalian) were selected from the TRRD database. All the promoter sequences were aligned with the TSS.

The Codon adaptation index (CAI) (Sharp, Li, 1987) was calculated by the CodonW 1.3 program (http://www.molbiol.ox.ac.uk/cu) and was used as a measure of the yeast gene expression level. The PHASE and RECON computer programs were used in comparative analysis of the nucleosome formation ability. The RECON program calculates the nucleosome formation potential (NFP) (Levitsky *et al.,* 2001). The PHASE program estimates the density of dinucleotides phased with helical turn periodicity in DNA sequences. The PHASE program yields the function PDD (periodic dinucleotide density) (Levitsky *et al.,* this issue). In the current study, we applied PDD only to AA and TT dinucleotides, smoothing window size was 145 bp.

Results

The profiles of the PDD and NFP for the promoter and 5'UTR regions of the yeast genes sequences aligned with the TSS is shown in Fig. 1. Clearly, the PDD is maximal in the wide [-100; +50] region overlapping the TSS. The NFP has, as a rule, constant values in the upstream region and slightly increased in the downstream.

The correlations between the PDD values and CAI and also between NFP values and CAI were determined (Fig. 2A, B). The correlation coefficients are mostly positive. Significance ($p < 0.05$) is observed in the [+5; +30] region for NFP, and the highest positive correlations in the [-90; -70] and [-20; +20] regions for PDD.

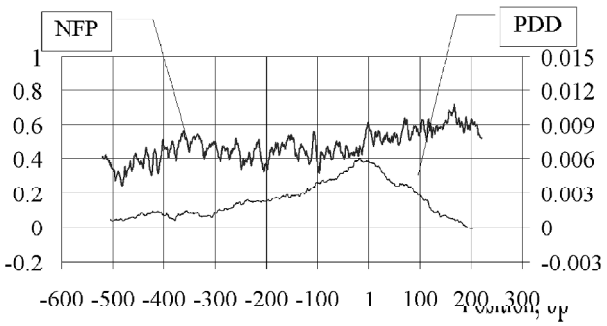


Fig. 1. Profile of periodic dinucleotide density (PDD) of AA and TT dinucleotides and nucleosome formation potential (NFP) for the promoter and 5'UTR regions of yeast genes.

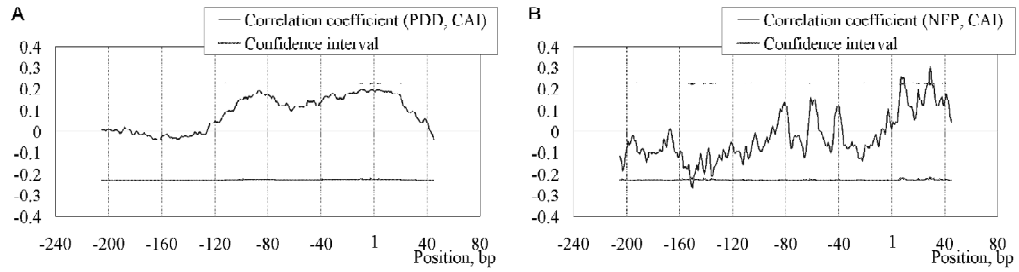


Fig. 2. Profile of the correlation coefficients between: (A) the periodic dinucleotide density (PDD) and CAI, (B) the nucleosome formation potential (NFP) and CAI for the promoter and downstream regions of the yeast genes.

Unlike yeast, the core promoters and downstream regions of *Mammalia* show other PDD and NFP profiles (Fig. 3): both profiles start to decrease about 400 bp upstream of the TSS and remain low within the downstream region.

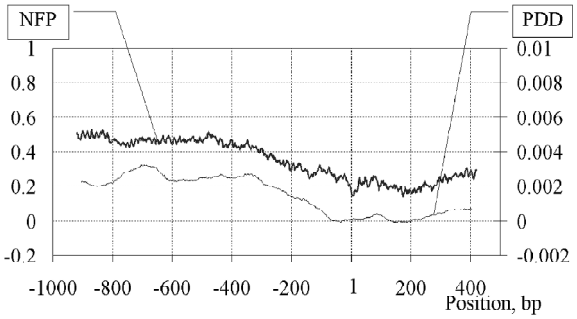


Fig. 3. Profile of the periodic dinucleotide density (PDD) and the nucleosome formation potential (NFP) of the TRRD gene sequences (*Mammalia*).

Discussion

It is known that dinucleotides AA and TT periodicities are most important in nucleosome site formation. Particular periodicities of these dinucleotides are related to the DNA curvature, thereby facilitating nucleosome formation (Ioshikhes *et al.,* 1996). The positive significant correlation between CAI and PDD for AA and TT dinucleotides found for the promoters may be related to DNA conformation just upstream of the TSS. This DNA conformation is very important in transcription initiation. The profiles of the correlation coefficients (Fig. 2A, B) along the yeast promoters and 5'-UTR regions revealed a wide range of positive correlations (from –100 to +40 bp). This may be interpreted as a tendency of the core promoters of the yeast highly expressed genes to form nucleosomes with a higher probability. It should be noted that the context parameters of the DNA nucleosome organization differ by the distribution pattern in the mammalian and yeast promoters.

Acknowledgements

The work was supported by the RFBR (grants Nos. 02-04-48802, 03-04-48555, 03-04-48829, 01-07-90376, 02-07-90355, 03-07-96833, 03-07-90181, 03-07-06078); Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); Russian Federal Research Development Program Research and Development in Priority Directions of Science and Technology (contract No. 38/2004); SB RAS (integration project No. 119); NATO (grants Nos. LST.CLG.979816), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology".

References

Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., Trifonov E.N. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences // J. Mol. Biol. 1996. V. 262. P. 129–139.

Kiyama R., Trifonov E.N. What positions nucleosomes? A model // FEBS Lett. 2002. V. 523. P. 7–11.

Kochetov A.V., Sarai A., Vorob'ev D.G., Kolchanov N.A. The context organization of functional regions in yeast genes with high-level expression // Mol. Biol. (Mosk.). 2002. V. 36. P. 1026–1034.

Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis // Bioinformatics. 2001. V. 17. P. 998–1010.

Levitsky V.G., Katokhin A.V., Furman D.P. A PHASE method analysis of dinucleotide content periodicities in nucleosomal DNA // This issue.

Lowary P.T., Widom J. Nucleosome packaging and nucleosome positioning of genomic DNA // Proc. Natl Acad. Sci. USA. 1997. V. 94. P. 1183–1188.

Pichueva A.G., Kochetov A.V., Milanesi L., Kondrakhin Yu.V., Kolchanov N.A. Correlations between sequence features of yeast genes functional regions and the level of expression // Bioinformatics of genome regulation and structure / Ed. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/ Dordrecht/London, 2004. P. 125–132.

Sharp P.M., Li W.H. The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications // Nucleic Acids Res. 1987. V. 15. P. 1281–95.

ANALYSIS OF PERIODICITIES IN THE DINUCLEOTIDE CONTEXT OF NUCLEOSOMAL DNA USING THE METHOD *PHASE*

Levitsky V.G.*, Katokhin A.V., Furman D.P.

Institute of Cytology & Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: levitsky@bionet.nsc.ru

Keywords: *nucleosome positioning, dinucleotide context periodicity, nucleosomal DNA*

Summary

Motivation: Accumulation of the data on the role of chromatin nucleosomal organization in transcription, replication, and other basic biological processes makes ever more important the development of software tools for detecting the patterns of nucleosome location in the genomic DNA.

Results: A method PHASE for assessing the density of phased dinucleotides in arbitrary DNA sequences was designed. The tool PHASE was used to analyze periodic characteristics of dinucleotide context of nucleosome formation sites (NFS). It was demonstrated that the density of periodic AA and TT dinucleotides along NFS was maximal in the regions encompassing positions ±41 relative to the site center, reflecting the highest capability of DNA regular bending there.

Availability: The software package realizing the tool PHASE is available at http://wwwmgs.bionet.nsc.ru/mgs/programs/phase/.

Introduction

Numerous data of experiments and computer studies suggest that trans-interactions of protein factors with their cognate sites and cis-interactions of the latter with nucleosome positioning sites determine an ordered location of nucleosomes in genomic DNA. The totality of these interactions forms the so-called code of nucleosomal chromatin organization (Lower, Widom, 1997; Trifonov, 1997; Kiyama, Trifonov, 2002).

Periodic location of di- and trinucleotides is among the most important characteristics of this code (Ioshikhes *et al*., 1996; Stein, Bina, 1999). Earlier, we demonstrated that particular characteristics of dinucleotide composition of nucleosomal DNA allowing the latter to attain the conformation necessary for DNA–histone interactions and form nucleosomes (Levitsky *et al*., 1999; 2001). It is also known that a periodic arrangement of AT-containing dinucleotides, in particular, phased location of AA and TT dinucleotides, make a most weighty contribution to the contextual code of nucleosome positioning (Ioshikhes *et al*., 1996; Kiyama, Trifonov, 2002). Thus, in this work, we focused our attention on analyzing these dinucleotide types and quantifying the periodicity in question by the tool PHASE, designed for this particular purpose.

Methods and Algorithms

A sample of 142 nucleotide sequences with a length of 300 bp carrying nucleosome formation site (NFS) in the center was used in the work. These sequences were extracted from EMBL databank entries according to the codes and positions indicated by Ioshikhes and Trifonov (1993).

The tool PHASE is based on estimation of the measure of similarity between the frequencies of phased dinucleotides in actual DNA sequences and the anticipated frequencies.

The dinucleotides located at a distance of one or two DNA helix turns (one turn on the average takes 10.5 bp) from each position considered were taken into account. Let us consider a nucleotide sequence S and a set of $N = 10$ periods $\{D_n\} = \{\pm10, \pm11, \pm20, \pm21, \pm22$ bp$\}$ (five periods in both directions from any position of the sequence that is located at least 22 bp from the sequence borders).

Let us calculate the complete set of dinucleotide frequencies in the sequence *S* and generate then a sample of randomized sequences {*R*} with the same dinucleotide composition. Let the dinucleotide of *j*th type ($1 \le j \le 16$) be present at the *i*th position of a sequence from the sample {*R*}. Let the number of dinucleotides of the same type as in this position that are located at positions $\{i+ D_n\}$ be *k* ($0 \le k \le N$). Let us designate the probability of observing *k* periodically located dinucleotides of *j*th type in all the sequences of sample {*R*} as $p_j(k)$. Then the frequency of dinucleotides of *j*th type, f_j, in the sample {*R*} is determined as the following sum:

$$f_j = \sum_{k=0}^{N} p_j(k). \quad (1)$$

Note that the probability $p_j(0)$ means the absence of periodic dinucleotides or the presence of exclusively "nonperiodic" dinucleotides of *j*th type, whereas the probability $p_j(N)$ corresponds to the highest periodicity of dinucleotides permitted by the set $\{D_n\}$.

Let us consider again the sequence *S* and assume that a dinucleotide of *j*th type is present at its *i*th position. Let *k* dinucleotides of *j*th type ($0 \le n \le 10$) be observed at positions $\{I + D_n\}$. Then, we determine the next function, $w_j(k)$, as a logarithm of the relative probability of occurrence of the *j*th type periodic dinucleotides:

$$w_j(k) = -\lg \frac{p_j(k)}{f_j}. \quad (2)$$

The anticipated mean value of the function w_j is equal to

$$w_j = -\sum_{k=0}^{N} \left(\frac{p_j(k)}{f_j} \lg \frac{p_j(k)}{f_j}\right). \quad (3)$$

The maximal value of the function w_j, as follows from definition (2), amounts to $w_j(N)$. Basing on the mean and maximal values of the function $w_j(k)$ (2), let us determine the function $PDD_j(k)$ of the periodic dinucleotide density (PDD) as follows:

$$PDD_j(k) = \frac{w_j(k) - w_j}{w_j(N) - w_j}. \quad (4)$$

Let us define the integral function PDD of the periodic densities for an arbitrary set of dinucleotides {*J*} as follows:

$$PDD = \sum_{j \in \{J\}} PDD_j. \quad (5)$$

As it follows from definition (4), the maximal value of functions (4) and (5) over the sequence positions equals unity, whereas the zero value corresponds to the mean calculated for randomized sequences, i.e., in the absence of phased dinucleotides.

Results and Discussion

The software package PHASE was used to analyze periodic contextual characteristics of sequences containing NFS. Fig. shows the profile of PDD function for AA and TT dinucleotides constructed for the sample of sequences containing NFS using an averaging window of 10 bp. The region [–73; +73] corresponds to NFS; regions [–110; –74] and [+74; +110], to linker regions. Analysis of the PDD

profile in the region [–73; +73] suggested the following inferences: (i) the maximal PDD values for AA and TT dinucleotides are observed in the regions encompassing positions ±41 and (ii) the minimal PDD values for the same dinucleotides are found in a wider regions surrounding positions ±89 and ±1.

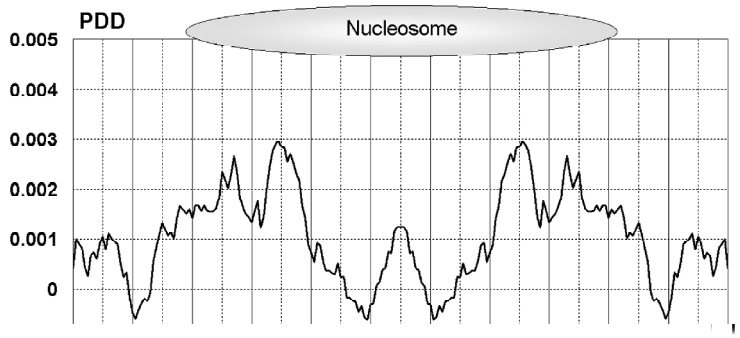


Fig. PDD profile (with an averaging window of 10 bp) for AA and TT dinucleotides for the sample of nucleotide sequences containing NFS (indicated by oval) in the center.

The linker regions also display low PDD values. These results comply well with earlier published data. For example, multiple alignment of nucleosome DNA sequences demonstrated that pronounced periodic characteristics of the context were absent in the central region of NFS (Ioshikhes *et al*., 1996). The result we obtained agrees as well with the X-ray structure analysis data showing that the terminal 10-bp sequences of nucleosomal DNA, adjacent to the linker regions, are unbend (Luger *et al*., 1997) and are bound least stably to the core histones (Polach, Widom, 1995).

As the high values of PDD profile obtained by PHASE indicate the presence of a phased dinucleotide at a distance of one or two DNA helix turns, while the presence of phased dinucleotide itself determines an increase in the ability of DNA to bend regularly, we may infer that nucleosome DNA contains two more pronouncedly bent regions with a length of 40–50 bp at positions [–73; –25] and [+25; +73] relative to NFS center. These two regions are separated with a less bent region of 30–50 bp located at [–25; +25]. Calculation of characteristics of nucleosomal DNA 3D trajectory (Fitzgerald *et al*., 1994) lead the authors to similar conclusions, as it was demonstrated that bending pattern consisted of repeating units of two 50–60 bp bending elements separated by a 20–30 bp region of a low curvature.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 03-04-48555-a and 03-07-96833); Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); Russian Federal Research Development Program «Research and Development in Priority Directions of Science and Technology» (contract No. 38/2004); and Program for Basis Research of the Presidium of the Russian Academy of Sciences (contract No. 10002-251/П-25/155-270/200404-082); Project No. 10.4 of the RAS Presidium Program «Molecular and Cellular Biology».

References

Fitzgerald D.J., Dryden G.L., Bronson E.C., Williams J.S., Anderson J.N. Conserved patterns of bending in satellite and nucleosome positioning DNA // J. Biol. Chem. 1994. V. 269. P. 21303–21314.

Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., Trifonov E.N. Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences // J. Mol. Biol. 1996. V. 262. P. 129–139.

Ioshikhes I., Trifonov E.N. Nucleosomal DNA sequence database // Nucleic Acids Res. 1993. V. 21. P. 4857–4859.

Kiyama R., Trifonov E.N. What positions nucleosomes? A model // FEBS Lett. 2002.V. 523. P. 7–11.

Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Kolchanov N.A. Nucleosomal DNA property database // Bioinformatics. 1999. V. 15. P. 582–592.

Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis // Bioinformatics. 2001. V. 17. P. 998–1010.

Lowary P.T., Widom J. Nucleosome packaging and nucleosome positioning of genomic DNA // Proc. Natl Acad. Sci. USA. 1997. V. 94. P. 1183–1188.

Luger K., Mader A.W., Richmond R.K., Sargent D.F., Richmond T.J. Crystal structure of the nucleosome core particle at 2.8 A resolution // Nature. 1997. V. 389. P. 251–260.

Polach K.J., Widom J. Mechanism of protein access to specific DNA sequences in chromatin: a dynamic equilibrium model for gene regulation // J. Mol. Biol. 1995. V. 254(2). P. 130–149.

Stein A., Bina M. A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment // Nucleic Acids Res. 1999. V. 27. P. 848–853.

Trifonov E.N. Genetic level of DNA sequences is determined by superposition of many codes // Mol. Biol. (Mosk.). 1997. V. 31. P. 759–767.

NUCLEOSOME FORMATION POTENTIAL OF THE GENE REGULATORY REGIONS

Levitsky V.G.*, Proscura A.P., Podkolodnaya O.A., Ignatieva E.V., Ananko E.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: levitsky@bionet.nsc.ru

Keywords: *transcription regulation, nucleosome positioning*

Summary

Motivation: Nucleosome organization of DNA plays the key role in transcription initiation. Exhaustive computer analysis of nucleosome formation potential of the regulatory gene regions appears to be of interest.

Results: Using the RECON program for constructing nucleosome formation potential (NFP) profiles, we analyzed 5 samples of promoters specific in gene expression pattern and 57 samples of transcription factor binding sites (TFBS) regions with flanks. It was shown that the NFP values for promoters are related to tissue- and stage-specificity of gene expression. The NFP pattern of the TFBS regions may also serve as a characteristic of the open chromatin structure.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/

Introduction

An essential feature of eukaryotic DNA is its packaging in chromatin structure. Today, nucleosomes are recognized as highly dynamic units through which the eukaryotic genome can be regulated (Khorasanizadeh, 2004). Hence, the regulated processes of condensation-decondensation of chromatin during the preferential activation of genes are of undeniable importance to gene function. Strong experimental evidence is accumulating for nucleosome positioning in the promoter regions of the eukaryotic genes being of great importance in their regulation of transcription (Goriely *et al*., 2003). Currently, there is a growing number of reports about the interaction of different transcription factors (TF) with nucleosomal DNA (Hsiao *et al.*, 2002). For example, precise positioning of the DNA double helix on the surface of the histone octamer precludes binding of NFI and Oct-1/OTF-1 to their cognate sequences (Truss *et al.*, 1995). The interaction of TF with DNA is a dynamic process. Thus, it was demonstrated that the equilibrium constants describing this interaction decrease progressively from either side of the nucleosomal DNA toward its pseudodyad (Anderson, Widom, 2000). In this way, the distributed nucleosome positioning signals may be crucial in determining chromatin architecture *in vivo* in the regulatory regions. The context code of nucleosome positioning (Kiyama, Trifonov, 2002) controls both the nucleosome packaging in discrete chromatin regions and the accessibility of TFBS.

Based on the different TF-nucleosome interactions, the TFs are subdivided into those capable of binding to the DNA sites within nucleosome (type 1) and those incapable to do so (type 2). Type 1 include factors GR (Li, Wrange, 1995), SP1, GAL4, USF (Chen *et al.*, 1994), type 2 includes TBP, HSF (Taylor *et al.*, 1991) and NF1 (Blomquist *et al.*, 1996). To provide the interaction of type 2 factors with DNA, nucleosome should be displaced, or, at least, its conformational state should be remodeled occasionally by type 1 factors. Our task is to study of the NFP of gene promoters with different expression patterns and genomic sequences containing TFBS.

Methods and Algorithms

To build the NFP, a method based on discriminant analysis and on calculation of dinucleotide frequencies in the local regions of nucleosomal sites was applied (Levitsky *et al.*, 2001; Levitsky,

2004). The NFP positive values agree with significant predictions of nucleosome formation. We have used the confidence level $\alpha = 0.95$ in analysis.

A total number 211 promoters 300 bp long ([–300; +1] relative to the transcription start) were analysed. Classification of promoters according to the gene expression patterns allowed us to distinguish 5 promoter classes (Table). The patterns were identified on the basis of the information stored in the TRRD and literature sources. The nucleosomal organization of the TFBS DNAs of 57 types was studied in the fragments of 160 bp long regions, including an experimentally characterized site occupying the central position. The sequences were extracted from the TRRD database. The TFBS samples contained from 10 to 280 sequences.

Table. Samples of the promoter gene regions

Description	Number of sequences
Genes controlling biosynthesis of steroid hormones	32
Interferon regulated genes	44
Genes regulating intracellular cholesterol level	20
Cell cycle regulated genes	75
Erythroid specific regulated genes	40

Implementation and Results

To further clarify the relation between the nucleosome positioning in promoters and gene expression, we analyzed the NFP profiles for genes of 5 classes showing different expression patterns (Fig. 1).

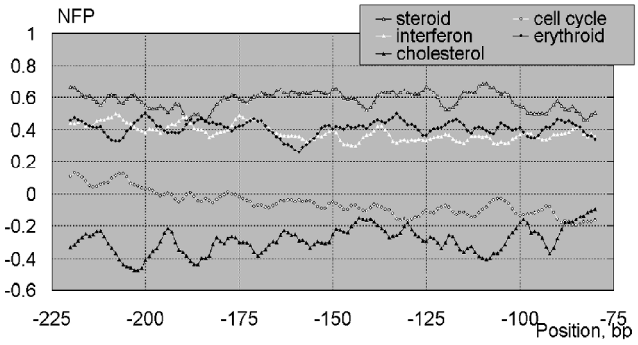


Fig. 1. Nucleosome formation potential profiles for the gene promoters showing different expression patterns: genes controlling biosynthesis of steroid hormones, interferon regulated genes, genes regulating intracellular cholesterol level, cell cycle regulated genes, erythroid specific regulated genes.

Figure 2 gives the NFP average values and confidence intervals for the sample of DNA fragments containing TFBS. It was found that most (46 of 57) of sites are located in the region with the positive NFP. Thus, all the sites were be subdivided into 3 groups: (i) sites with high positive NFP ($\varphi(X) > 0.5$, the total number of sites was 16); (ii) sites with low positive NFP ($0 < \varphi(X) < 0.5$, their number was 30); (iii) sites with negative NFP ($\varphi(X) < 0$, the total number of sites was 11).

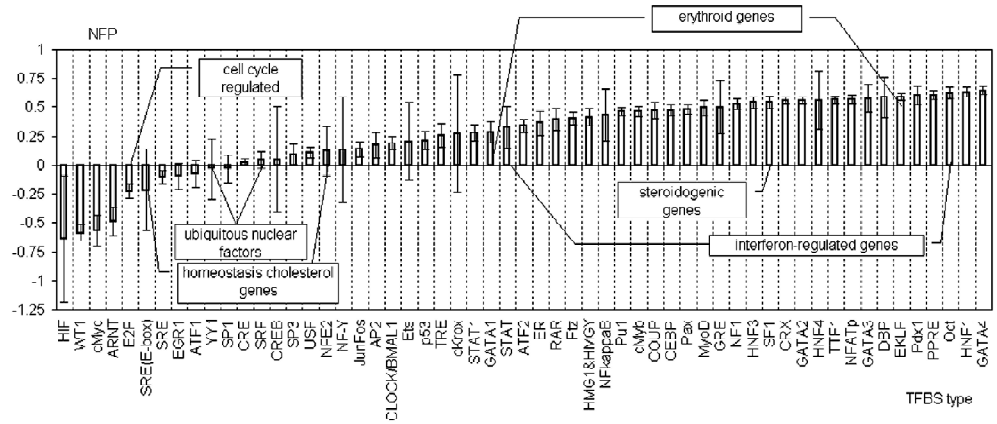


Fig. 2. Average values for nucleosome formation potential and respective confidence intervals for the sample of DNA fragments containing TFBS at the central position.

Discussion

Let us examine the NFP profile values for the promoters showing different expression patterns. The low values for the NFP of the genes regulating intracellular cholesterol level and cell cycle regulated genes presumably reflect the possible easy access of TF to the proximal promoters of these genes, which are regulated for rapid transcription initiation and providing high expression level. Let us now turn to the promoters of the erythroid specific regulated genes, genes controlling biosynthesis of steroid hormones and interferon regulated genes for which the NFP values are high. It should be noted that these genes are tissue-specific and inducible, in contrast to genes regulating intracellular cholesterol level and cell cycle regulated genes. For such genes nucleosome packaging in promoter DNA may be an important element in the regulatory mechanism of transcription. It may be assumed that the state of tightly packed chromatin is normal in this case. This state may be overcome by the appearance of appropriate TFs in cell nuclei. They interact with chromatin and alter or abolish DNA nucleosome packaging. Therefore, tissue-specific and inducible genes possess finer mechanisms for transcription initiation. Change in the nucleosome packaging pattern in promoter DNA and it transition from repressed to active state are indispensable for the transcription initiation mechanism.

Let us examine how the average NFP values for the sample of DNA fragments containing TFBS are distributed. It is noticeable that the distribution (Fig. 2) is consistent the reference of these site to a particular expression patterns of genes (Fig. 1). Thus, for example, the lowest NFP values were for the TFBS occurring in the cell cycle regulated genes (-0.23 E2F, -0.57 cMyc), the genes regulating lipid metabolism (-0.10 SRE), and also ubiquitous TFBS (-0.04 YY1, -0.03 SP1). The highest NFP values were for the TFBS most frequently occurring in the tissue-specific and inducible genes (0.62 Oct, 0.54 SF1).

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grants Nos. 02-04-48802, 03-04-48555, 03-04-48829, 01-07-90376, 02-07-90355, 03-07-96833, 03-07-90181); Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); Russian Federal Research Development Program Research and Development in Priority Directions of Science and Technology (contract No. 38/2004); NATO (grants Nos. LST.CLG.979816 and LST.CLG 979815), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”. The authors are

grateful to Drs. Khlebodarova T.M., Lokhova I.V., Mischenko E.L., Suslov V.V., Stepanenko I.L., Turnaev I.I., Orlov Yu.L. for supporting the work and to Prof. N.A. Kolchanov for valuable discussion.

References

Anderson J.D., Widom J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites // J. Mol. Biol. 2000. V. 296(4). P. 979–987.

Blomquist P., Li Q., Wrange O. The affinity of nuclear factor 1 for its DNA site is drastically reduced by nucleosome organization irrespective of its rotational or translational position // J. Biol. Chem. 1996. V. 271. P. 153–159.

Chen H., Li B., Workman J.L. A histone-binding protein, nucleoplasmin, stimulates transcription factor binding to nucleosomes and factor-induced nucleosome disassembly // EMBO J. 1994. V. 13. P. 380–390.

Goriely S., Demonte D., Nizet S., De Wit D., Willems F., Goldman M., Van Lint C. Human IL-12(p35) gene activation involves selective remodeling of a single nucleosome within a region of the promoter containing critical Sp1-binding sites // Blood. 2003. V. 101(12). P. 4894–4902.

Hsiao P.W., Deroo B.J., Archer T.K. Chromatin remodeling and tissue-selective responses of nuclear hormone receptors // Biochem Cell Biol. 2002. V. 80(3). P. 343–51.

Khorasanizadeh S. The nucleosome: from genomic organization to genomic regulation // Cell. 2004. V. 116(2). P. 259–272.

Kiyama R., Trifonov E.N. What positions nucleosomes? A model // FEBS Lett. 2002. V. 523. P. 7–11.

Levitsky V.G. RECON: a program for prediction of nucleosome formation potential // Nucleic Acids Res. 2004, (in press).

Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis // Bioinformatics. 2001. V. 17. P. 998–1010.

Li Q., Wrange O. Accessibility of a glucocorticoid response element in a nucleosome depends on its rotational positioning // Mol. Cell. Biol. 1995. V. 15. P. 4375–4384.

Truss M., Candau R., Chavez S., Beato M. Transcriptional control by steroid hormones: the role of chromatin. 1995. V. 191. P. 7–23.

DISTANCE PREFERENCES IN DISTRIBUTION OF BINDING MOTIFS AND HIERARCHICAL LEVELS IN ORGANIZATION OF TRANSCRIPTION REGULATORY INFORMATION

***Makeev V.J.*[1]*, *Lifanov A.P.*[2], *Nazina A.G.*[3], *Papatsenko D.A.*[3]**

[1] Scientific Center "Genetika", Moscow, Russia; [2] Engelhardt Institute of Molecular Biology, Moscow, Russia; [3] Department of Biology, New York University, New York, USA
* Corresponding author: e-mail: makeev@imb.ac.ru

Keywords: *transcription, regulation, factor, binding site, enhancer, periodicity, composite element*

Resume

Motivation: Initiation of eukaryotic transcription is a complex process that involves huge protein complexes, which interact with regulatory DNA at many DNA-protein binding sites. A huge amount of regulatory information is contained in the mutual positioning of binding sites. Investigation in detail of interrelation of site location can help to understand the spatial conformation of DNA helix within these regions as well as facilitate developing of new computerized methods of locating of regulatory modules.

Results: We explored distance preferences in distribution of binding motifs for five transcription factors (Bicoid, Kruppel, Hunchback, Knirps and Caudal) in a large set of *Drosophila Cis*-Regulatory Modules (CRM). We established that the vast majority of high-affinity binding sites are positioned in the CRMs at distances close to 10, 20, 30 etc. base pairs, i.e. approximately on the same side of the DNA helix. We also assess site overlapping and positioning of site pairs at specific distances. We discuss hierarchical levels in organization of transcription regulatory information and its role in detection of transcription regulatory regions in genome.

Availability: http:/homepages.nyu.edu/~dap5/

Introduction

Initiation of tissue-specific or spatio-specific transcription in multicellular organisms requires binding of multiple transcription factor molecules to transcription regulatory regions, such as promoters and enhancers. Multiplicity of binding motifs and binding sites for the same motif in the regulatory regions, are often characterized as regulatory clusters (Wagner, 1999). However, the number of sites and their affinity is not sufficient to describe regulatory information encoded by the binding motifs: specific *arrangement* of these binding motifs within the regulatory regions appears to be necessary for achieving proper biological function. The arrangement of binding motifs carries signature of 3D protein complexes, involved in the initiation of specific transcription. The precisely arranged pair of distinct binding motifs, involved into formation of specific DNA-protein-protein complexes are called composite elements (CE). The database of composite elements, TRANSCompel (Kel-Margoulis *et al*., 2002) combines 256 (Release 6.0) elements from different organisms.

In the current work we explored presence of preferential site distances in *cis*-regulatory modules (CRM) of *Drosophila* developmental genes. The term CRM stands for extended transcription regulatory units (~1Kb range), often located far from transcription start site and responsible for spatio-temporal expression of their cognate developmental genes (Berman *et al*., 2002). Database of known functional *Drosophila* CRMs is available (see NYU web site: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm) along with the list of matrices for a number of transcription factors and known transcriptional interactions, thus providing an excellent original dataset for computational study.

Results

Measuring distances between binding sites requires careful mapping of binding motifs in a relevant set of sequence data. We selected Position Weight Matrices (PWMs), generated from good alignments and limited our choice by five best known binding motifs for transcriptional regulators Bicoid, Caudal, Hunchback, Kruppel and Knirps, having relatively large number of instances in our CRM database. At the next step, we selected from our database of experimental CRMs only sequences containing significant homotypic clusters for these five binding motifs (Lifanov *et al*., 2003).

The total size of analyzed CRMs after the described prescreening procedure comprised more than 68 Kb of sequence data and contained 58 separate homotypic clusters in 33 non-overlapping contigs (see: http://homepages.nyu.edu/~dap5/PCL/appendix2.htm) along with the contigs themselves (Papatsenko *et al*., 2002; Lifanov *et al*., 2003).

We measured the distances between the centers of binding site cores, which were established on the basis of informational content of motif. We considered motifs at the same or at different DNA strands, corresponding to the tandem of motif arrangement respectively.

First we compared the distribution of binding motifs in a hole set of regulated CRMs using the geometric distribution evaluated from the same set as the reference hypothesis. The comparison was performed for different levels of site quality, i.e. different PWM cutoff value. To insure site independence we excluded close distances (less than 5 bases), which often correspond to the overlapping sites. The described distance distribution test clearly shows that the binding sites in Drosophila CRMs are distributed in a non-random fashion and large fraction of them has spacing not exceeding 50–60bp (see (Makeev *et al*., 2003)). This distance range is comparable with the size of composite elements (CE) reported earlier.

To test whether the binding motifs distributed in the CRMs periodically we calculated all distances between the binding sites. In a random Bernoulli sequence the all-distance occurrence has uniform distribution, independent from the distance itself. Distance distribution between particular different binding sites may be also interesting, especially if they involved in synergistic or antagonistic interactions. Fourier spectra were used to assess periodical positioning.

First, we built distributions and the corresponding Fourier spectra for motifs of the same types. The most striking result, entirely confirming the hypothesis of 'helical phasing' was obtained for Bicoid. The vast majority of the high-affinity Bicoid sites are positioned at distances close to 10, 20, 30 etc. base pairs (see the Figure). The periodicity in the distribution of Bicoid (Bcd) sites drops very fast with decreasing site affinity (PWM score), supporting biological role of this specific spacing, rather than short-range correlations in DNA sequence itself.

Similar, but not identical periodic signal was revealed in the distribution of Hunchback (Hb) (see the Fig.). In this case however, the period was equal not to 10 but to 11 base pairs. We believe that this difference in periodicity is the result of slightly different DNA conformation of the two binding motifs (compare CCTAATCCC, – consensus for Bicoid and TTTTTTTG, – consensus for Hunchback). Surprisingly, but distribution of another binding motif, Kruppel (Kr), showed no presence of periodic signals, corresponding to the 'helical phasing'. This can be a reflection of different biological role of those factors. Whereas Bicoid and Hunchback are transcriptional activators mostly, involved in cooperative polymerization and protein-protein interactions with other transcription factors, Kruppel is a repressor protein. Caudal motif alone also displayed 'helical phasing' and it is also an activator protein. Knirps motif had low number of occurrences in our dataset and was not considered alone.

The analysis of distance preferences between specific combinations of binding motifs demonstrated that Bcd-Hb hunchback motif pair demonstrated lower periodic signal than the sites for both proteins taken independently. In contrast the motif pair, Bcd-Kr shows a prominent periodic signal the phase of which contrasts the helical pitch (17 bp) (Makeev *et al*., 2003). Thus, non-

overlapping Bcd and Kr sites have a tendency to be distributed on opposite surfaces of DNA. This result may suggest that the non-overlapping Bcd and Kr sites may belong to distinct composite elements, performing different functions.

Finally, we extracted periodic signals from combination of all five binding motifs and detected presence of the same 'helical phase' signal, though having smaller amplitude and, likely, contributed by distances Bcd-Bcd and Hb-Hb. It is important, however, that the distribution of non-overlapping binding motifs in regulatory regions cannot be described by the simple 'helical phasing' formula. The more detailed description of these finding can be found in (Makeev *et al*., 2003).

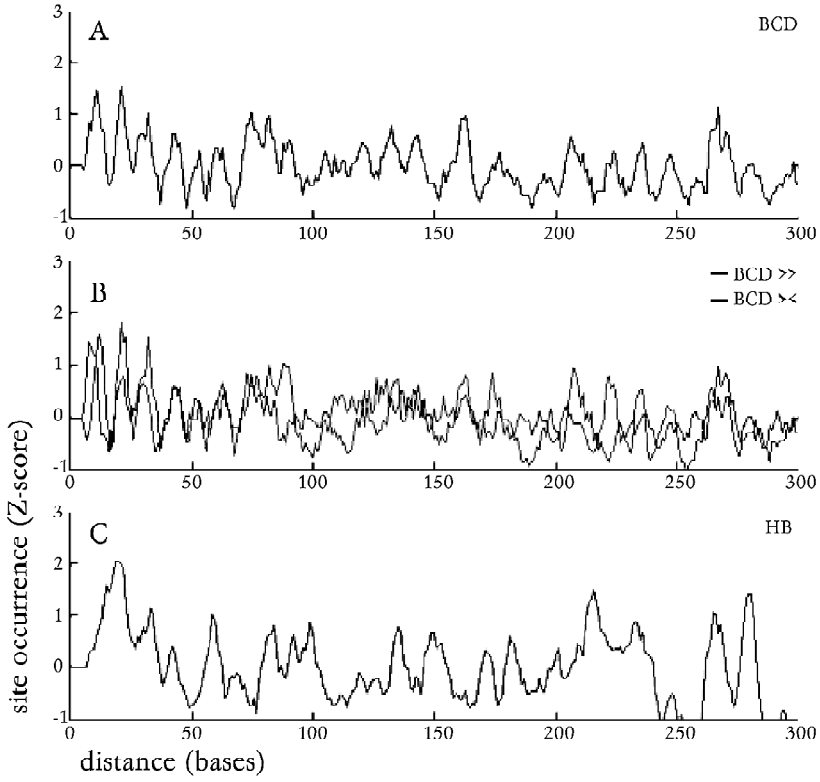


Fig. All-distance distribution for Bicoid and Hunchback.
Panel A shows Z-scores calculated for distribution of distances found between any two Bicoid sites. On the panel B the same signal is split for Bicoid matches found in tandem (>>) or palindrome (><) orientation. Panel C shows distance distribution for Hunchback sites.

Discussion

Binding motifs are distributed in Drosophila CRMs in a non-random fashion, where a large fraction of sites exhibit distances not exceeding 50–70 base pairs. We believe that the composite elements represented by small groups (2–5) of specifically arranged (spaced) binding motifs comprise the intermediate organizational level. Size of promoter or cis-regulatory module (05–1Kb) allows fitting several such composite elements, perhaps acting independently in their response to verity of transcriptional signals. For example, repression of the same promoter (CRM) by two transcription factors might be achieved trough independent action of two corresponding composite elements. In this respect, our finding of the opposite to the helical phase (x17 bp) in distribution of Bcd-Kr pair represents special interest.

Acknowledgements

This study was partially supported by RFBR grant 04-07-90270-в, and by Program in Molecular and Cellular Biology of Russian Academy of Sciences, coordinator V.G. Tumanyan, V. Makeev and A. Lifanov are also supported by CRDF RBO-1268-MO-02 and Howard Hughes Institute Grant 55000309.

References

Berman B.P., Nubi Y., Pfeiffer B.D., Tomancak P., Celniker S.E., Levine M., Rubin G.M., Eisen M.B. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in Drosophila genome // Proc. Natl Acad. Sci. USA. 2002. V. 99. P. 757–762.

Kel-Margoulis O.V., Kel A.E., Reuter I., Deineko I.V., Wingender E. TRANSCompel: a database on composite regulatory elements in eukaryotic genes // Nucleic Acids Res. 2002. V. 30. P. 332–334.

Lifanov A.P., Makeev V.J., Nazina A.G., Papatsenko D.A. Homotypic regulatory clusters in Drosophila // Genome Res. 2003. V. 13. P. 579–588.

Makeev V.J., Lifanov A.P., Nazina A.G., Papatsenko D.A. Distance preferences in the arrangements of binding sites and hierarchical levels in organization of transcription regulatory information // Nucleic Acids Res. 2003. V. 31. P. 6016–6026.

Papatsenko D.A., Makeev V.J., Lifanov A.P., Regnier M., Nazina A.G., Desplan C. Extraction of functional binding sites from unique regulatory regions: the Drosophila early developmental enhancers // Genome Res. 2002. V. 12. P. 470–481.

Wagner. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes // Bioinformatics. 1999. V. 15. P. 776–784.

EXTREMELY CONSERVED NON-CODING SEQUENCES IN VERTEBRATE GENOMES

***Makunin I.V.*[1], *Stephen S.*[1], *Pheasant M.*[1], *Bejerano G.*[2], *Kent J.W.*[2], *Haussler H.*[3], *Mattick J.S.*[1]**

[1] ARC Special Research Centre for Functional and Applied Genomics, Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD 4072, Australia; [2] Department of Biomolecular Engineering, University of California Santa Cruz, Santa Cruz, CA 95064, USA; [3] Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, CA 95064, USA

***** Corresponding author: e-mail: i.makunin@imb.uq.edu.au

Keywords: *ultra-conserved sequences, comparative vertebrate genomics*

Summary

Motivation: We are exploring the thesis that the majority of the human genome, and that of other complex organisms, encodes an expanded regulatory system primarily transacted by RNA, and that much of the genome is under evolutionary selection, both positive and negative. Therefore we have been examining the global conservation patterns between genomes, concentrating initially on those sequences that are most highly conserved.

Results: There are 741 sequences over length 200 bp identical between human and mouse genomes, and 481 sequences over 200 bp are identical between human, mouse and rat. Approximately half of them, 222, are located in intergenic regions, and the rest reside within genes. Ultra-conservative elements tend to be in clusters on chromosomes and situated within or close to regulatory genes. Ultra-conserved sequences are present in chicken with average level of identity about 95 % and dog at 99 %, and many can be traced back to amphibians and fish. This level of conservation suggests a substitution rate of more than one order of magnitude lower than for protein coding regions. We speculate that the extreme conservation of at least some of these sequences is determined by selection at the RNA level.

Introduction

Analysis of available rodents and human genomes has suggested that the fraction of the genome which is conserved above neutral evolution is about 5 %, considerably in excess of that accounted for by protein-coding sequences, about 1.2 %. Indeed, recent reports indicate that some non-coding sequences in vertebrate genomes are more conserved than protein coding regions (Dermitzakis *et al*., 2003). Here we describe an unusual group of extremely highly conserved sequences within vertebrate genomes. These long sequences remain almost identical in mammals and can be traced back to fish.

Methods and Algorithms

Methods and algorithms are described in Bejerano *et al*. (2004).

Implementation and Results

We designed a search for extremely highly conserved, namely long identical, sequences between human-mouse, human-rat or human-mouse-rat genomic alignments from UCSC database (http://genome.ucsc.edu/). 741 identical sequences equal or over 200 bp were found in human-mouse alignment, and most of them, except for 28 (approximately 4 %), are highly conserved in rat genome. We used the following criteria for conservation: the presence of at least one 25 bp-long identical sequence in the corresponding region of the human-rat alignment. Comparison with a

third species allowed us to exclude possible cross-species contamination during genome assembly. There are 675 identical sequences equal to or over 200 bp present in the human-rat alignment, and only 7 of them (approximately 1 %) are not highly conserved in the mouse genome. It seems that the observed difference in the number of ultra-conserved elements absent in the third species is due to the quality of sequence in the mouse and rat genomes. Finally, 481 identical sequences equal or over 200 bp are present in human, mouse and rat genomes. Among them, only 111 overlap annotated exons, within 93 genes, and the rest are present in introns or intergenic regions. This group is enriched in genes involved in regulation of splicing and proteins containing the RNA recognition motif (for example, SFRS3, SFRS11, HNRPU and some others). Genes containing ultra-conserved elements within introns or in adjacent regions are enriched in DNA binding proteins and transcription regulators (Bejerano *et al*., 2004). This is just tip of the iceberg as there are more than 5.000 sequences of over 100 bp that are identical between human, mouse and rat.

The presence of identical sequences in organisms separated more than 85 million years (Myr) ago indicates an extremely high level of conservation. We compared human ultra-conserved elements with other vertebrates: dog, chicken, frog, and pufferfish sequences. The 477 sequences from 481 are present within dog traces at an average of 99.2 % identity. The 467 out of 481 elements align to chicken genome at an average of 95.7 % identity. The estimated mutation rate for some of these elements between human and chicken is 0.00004 per nucleotide per million years. That is one order of magnitude lower that mutation rate for most conserved non-coding sequences described in Dermitzakis *et al*. (2003) and 20 times lower than for protein-coding sequences. Some of the ultra-conserved elements could be traced to pufferfish although the extent and level of identity is much smaller than within mammals, birds and amphibians. One extreme example of conservation are the intronic ultra-conserved elements in the POLA gene, one of which possesses up to 90 % identity over a 500 bp sequence to pufferfish and zebrafish.

Ultra-conserved elements tend to be clustered on chromosomes (Fig.). This is most obvious for long sequences. For example, on X-chromosome all but two ultra-conserved sequences over 300 bp are located within or close to the POLA gene (Fig.). Other examples of clustering are adjacent to the FOX2 gene, which is mutated in a severe speech and language disorder (51 identical sequences over 100 bp between human and mouse), and the EBF3 gene, which is involved in neuronal differentiation and regional specification in the central nervous system (44 identical sequences over 100 bp between human and mouse). Long chromosomes tend to have higher number of ultra-conserved elements, the highest being 110 identical human-mouse sequences on the second chromosome. Only a single sequence over 200 bp, identical between human and mouse, is present on chromosome 21, and no long ultra-conserved sequences were found on the Y chromosome.

Discussion

The observed conservation, namely long identical sequences, can not be explained by protein coding constraints (no third base periodicity) or by conservation based on protein-DNA interaction (usually proteins bind to very short sequences). One might speculate that some of the ultra-conserved elements could be distal enhancers (Nobrega *et al*., 2003) but this does not tell a lot about their possible mechanism of action or the basis of their strong conservation. Several observations indicate that at least in some cases the observed conservation might depend on selection at RNA level. First, ultra-conserved elements often overlap splicing sites of alternative exons or present in regions subjected to RNA editing, such as GRIA3 gene (Aruscavage, Bass, 2000). Second, many ultra-conserved elements are present in genes encoding protein involved in splicing regulation: splicing factors (SFRS genes), heterogeneous nuclear ribonucleoprotein (HNRP) genes (which are antagonistic to SFRSs), and some others. In these genes ultra-conserved elements are present in regions possessing evidence for alternative splicing and intron retention.

Moreover, there is strong evidence for alternative splicing and intron retention for both human and mouse sequences, and in some cases for other species including chicken. Splicing and intron retention depends on RNA rather than on DNA sequences. We speculate that observed ultra-conservation in some cases might be associated with RNA regulation, as has been proposed previously (Mattick, 2004).

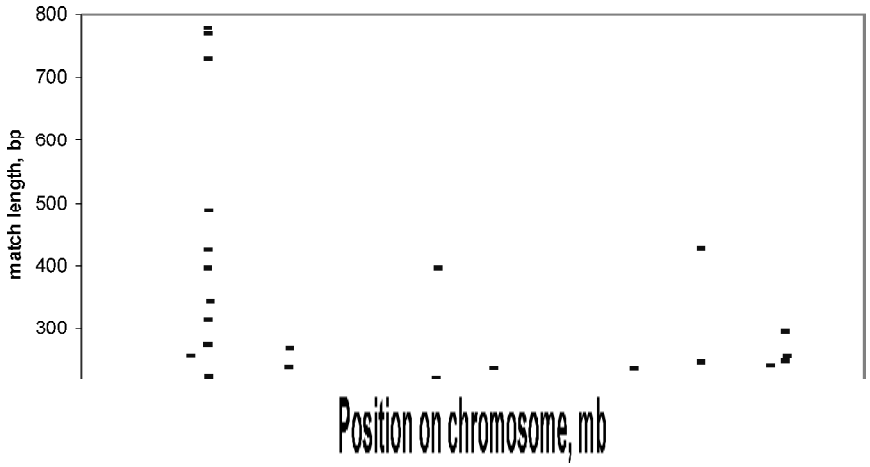


Fig. Distribution of ultra-conserved elements on human X-chromosome. Length of 40 identical human-mouse sequences (Y) was plotted against position on X-chromosome (X). The square corresponds to the centromeric region, the asterisk indicates the position of the POLA gene.

Acknowledgements

S.S., M.P., I.M. and J.S.M. were supported by the Australian Research Council and the Queensland State Government. G.B., W.J.K. and D.H. were supported by NHGRI grant 1P41HG02371, NCI contract 22XS013A, and D.H. additionally by the Howard Hughes Medical Institute.

References

Aruscavage P.J., Bass B.L. A phylogenetic analysis reveals an unusual sequence conservation within introns involved in RNA editing // RNA. 2000. V. 6(2). P. 257–69.

Bejerano G., Pheasant M., Makunin I., Stephen S., Kent J.W, Mattick J.S., Haussler H. Ultra-conserved elements in the human genome // Science. Published online 6 May 2004. [DOI: 10.1126/science.1098119].

Dermitzakis E.T., Reymond A., Scamuffa N., Ucla C., Kirkness E., Rossier C., Antonarakis S.E. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs) // Science. 2003. V. 302(5647). P. 1033–5.

Mattick J.S. RNA regulation: a new genetics? // Nature Rev. Genetics. 2004. V. 5. P. 316–23.

Nobrega M.A., Ovcharenko I., Afzal V., Rubin E. M. Scanning human gene deserts for long-range enhancers // Science. 2003. V. 302(5644). P. 413.

COMPUTER-BASED ANALYSIS AND RECOGNITION OF POTENTIAL IRON-RESPONSIVE ELEMENTS IN 5' AND 3' UTR TRANSCRIPTS OF EUKARYOTIC GENES

Mishchenko E.L.*, Kondrakhin Yu.V., Podkolodnaya O.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: elmish@bionet.nsc.ru

Keywords: *iron metabolism, RNA elements, iron-responsive elements, iron regulatory proteins, cellular iron homeostasis, posttranscriptional gene regulation*

Summery

Motivation: Interaction of iron regulatory proteins with iron-responsive elements located in 5'- and 3'UTR mRNA is one of the fundamental mechanisms of regulation of gene expression acting at the posttranscriptional level. Currently, only small amount of these elements were revealed experimentally. So, the study presented is devoted to recognition of potential iron-responsive elements in genome sequences, as well as to their computer-based analysis.

Results: We have analyzed all rodent mRNA-sequences stored in GeneBank and supplied with CDS annotation. The total length of the sequences analyzed equals approximately to 80,000,000 bp. We have detected all functional IREs described for this group of organisms. Statistically significant (p-value <0,01) increase in IRE-sites occurrence in non-translated mRNA regions was found both in 5'UTR and 3'UTR. Saturation by sites was essentially higher in 5'UTR than in 3'UTR. Probability of a site occurrence in the coding regions coincides to the random probability. Analysis of nucleotide content of potential sites has revealed that the most and lest conserved are the 10-th and 11-th site positions, respectively.

Introduction

Iron-responsive elements (IREs) are regulatory elements located in 5'- and 3'UTR mRNA and important for translation regulation of genes controlling iron homeostasis. Stem-loop structures of IREs specifically bind two types of iron-regulated proteins, IRP 1 and IRP2. Functional properties of these proteins are dependable upon iron concentration in a cell. Currently, two main types of stem-loop structures, of type I and type II, are revealed (Fig. 1a,b). IRE of type I is described in 5'UTR of ferritin gene and is the most frequently occurring (Haile, 1999). Functional IREs were also found in 5'- and 3'UTR mRNA of vertebrates, insects, bacteria, and protozoa, thus, evidencing about evolutionary conservatism of this mechanism of gene expression regulation.

Highly conserved stem-loop structure of IRE type I is characterized by the hexanucleotide loop with consensus CAGYGX, where Y=U,A, and X=U,C,A. Nucleotides C1 and G5 form a complementary pair. Adjacent to the loop upper stem consisting of five pairs of complementary bases is separated from the lower stem of variable length by unpaired cytosine (Fig. 1a) (Addess *et al*., 1997).

Under conditions of inner cellular iron depletion, cytosolic iron regulatory proteins, IRP1 and IRP2, are able to interact specifically and effectively with IREs of both types (Kd = 30–50 pM for IRE(I)•IRP2 and IRE(I)•IRP1, respectively) (Rogers *et al*., 2002; Butt *et al*., 1996]. The resulting IRE•IRP complexes in 5' UTR mRNA suppress translation, by blocking association

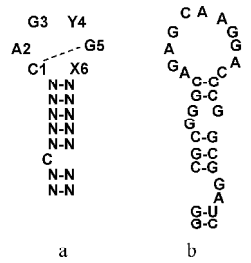


Fig. 1. Variants of IRE's stem-loop structures; a) type I; b) type II; Y=U or A; X=U, C, A.

of the 43S preinitiation translation complex with mRNA. IRE•IRP complexes in 3'UTR mRNA stabilize transcript and promote effective translation. Dissociation of IRE•IRP complexes takes place under inner cellular iron saturation. On these conditions, IRP1 coordinates [4Fe-4S] cluster and reveals aconitase activity, while IRP2 becomes a target for degradation by proteosomes (Haile, 1999).

IRPs may effectively, but with less affinity, bind to some mutated IRE-targets of type I. So, for effective interaction of IRP1 with IRE in H-ferritin, the loop substitutions of Ñ1G5 for U1A5 or G1C5 (Fig. 1a), which also form a complementary pair, are possible. However, for interaction with IRP2, only substitution of C1G5 for G1C5 is possible. Neither of the proteins may bind under conditions of substitution of C1G5 for A1U5. The substitutions of the loop nucleotides N2, N3, and N4, participating in direct contact with IRPs, cause formation of overlapping, but differing for IRP1 and IRP2 targets (Henderson *et al*., 1996). As shown, some point mutations, affecting both loop nucleotides and nucleotides of the upper and lower stems and breaking the A-form of their spiral conformation, may have a strong influence on affinity of IRPs to IRE of the human L-ferritin. Thus, they cause hyper-expression of L-ferritin and, finally, hereditary hyperferritinemia-cataract syndrome (Allerson *et al*., 1999).

Thus, interaction IRE•IRP may be viewed as the example of the fundamental mechanism of gene expression regulation acting at the posttranscriptional level. During this process, IRP1 and IRP2 are the functional sensors of iron homeostasis in a cell. However, up to nowadays, only 15 functional natural IREs of type I were experimentally revealed. The goal of the present study was to make a computer analysis and predict potential IREs of the type I in genome sequences. Since IRE type II is rarely occurring variant of IRE, it was not taken into analysis in the study presented here.

Methods

For searching for potential IRE-sites in nucleotide sequences, we have applied the following rules. The total length of a site was accepted as 19 bp. The second position of a site contains unpaired cytosine separating upper and lower stems of the loop structure. The first and 19-th positions of a site are represented by the only complementary base pair composing the lower stem. The upper stem is represented by five pairs of complementary positions located in 3–7 and 14–18 positions of a site simultaneously. Finally, the loop has a consensus CAGUGN and at that we have considered two variants: perfect correspondence to consensus and I perfect correspondence admitting one mismatch. Under description of all complementary pairs composing the stem, we accept interaction U-G in addition to C-G and A-U interactions.

Results and Discussion

By using the scheme described and aimed at recognition of the IRE-site, we have analyzed all rodent mRNA-sequences stored in GeneBank entries with CDS annotations. The total length of all analyzed sequences was approximately 80000000bp. It should be noted that we have revealed all functional IREs, described for this group of animals. The histograms of the relative frequencies of occurrence of potential IRE-sites in various mRNA regions (i.e., 5'UTR, CDS, and 3'UTR) are given in Figure 2. Additionally, the last columns of histograms correspond to the values of theoretical probability of finding the site in the random nucleotide sequences with equally occurring nucleotides. Statistically significant (p-value <0,01) increase of occurrence of IRE-sites in non-translated mRNA regions was found under considering both variants of the loop (Fig. 2a; perfect correspondence with consensus; Fig. 2b, one admissible substitution in a loop). Notably, saturation by sites in 5'UTR was essentially higher than that of 3'UTR. On the other hand, probability of the site occurrence in the coding regions coincides to the accidental probability.

We have analyzed the nucleotide content for all 19 positions of potential sites found in 5'UTR under admission of one mismatch. In Figure 3, one may see the entropy values illustrating the variability level of nucleotide content at each position of the site.

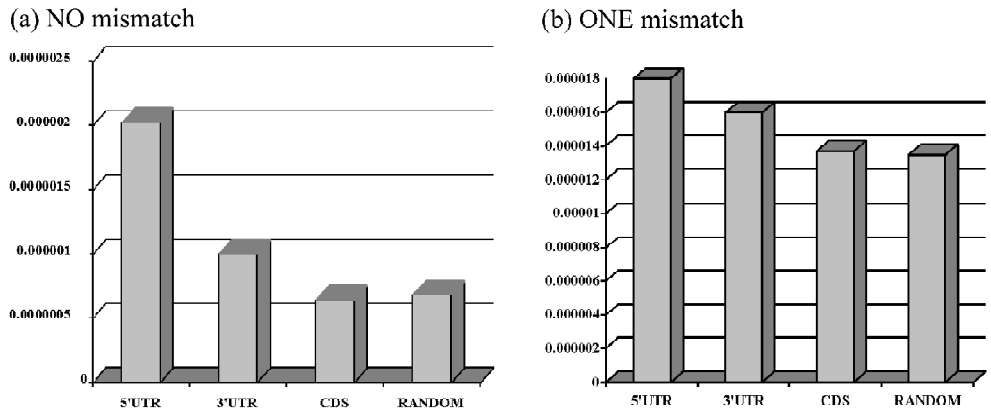


Fig. 2. Histogram of relative occurrence frequencies (probabilities) of potential sites in different regions of mRNA (5'UTR, 3'UTR and CDS), under variable number of mismatches in a conservative region of a loop: a) no mismatch; b) one mismatch.

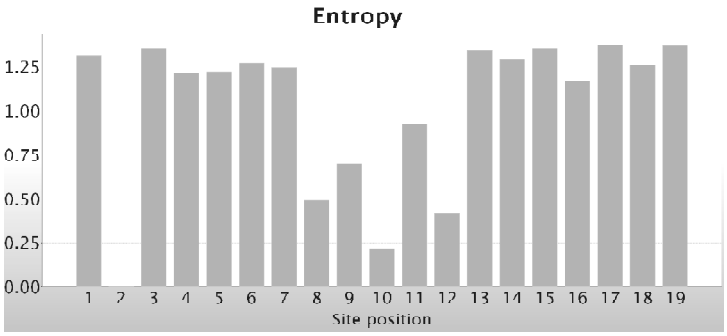


Fig. 3. Entropy values for all 19 positions of potential sites found in 5'UTR under condition of a single mismatch.

As was expected, in all positions of the site compiling complementary pairs of a loop, we have found no preferences for this or that nucleotide. For five positions (ranging from 8-th to 12-th positions) constituting the loop, we have observed some irregularity in the entropy values. This fact evidences to different level of evolutionary conservation in these positions and correlates well with experimental data. In particular, the most conservative is the 10-th position of the site. Mutational substitutions in this position cause, as a rule, the loss in functionality of the whole site (Allerson *et al*., 1999). At the same time, the neighboring 11-th position is characterized by the most variability of nucleotide content in comparison to 5 positions conside-red. This fact is also in a good agreement with the data on functional site structure published in literature (Henderson *et al*., 1994).

More detailed description of the nucleotide content in all 5 positions is represented in Table in a form of frequency matrix. It should be noted that the frequency matrix presented is similar to the matrix obtained by in vitro selection of IRE type I (Henderson *et al*., 1994). We suppose that by using this matrix instead of consensus CAGUG, it is possible to increase reliability of results of recognition of potential IRE-sites. Further analysis of the loop structure of the IRE type I, as well as the studying of IRE type II, are necessary for constructing full-value method aimed at searching for potential IRE.

Table. Nucleotide frequency matrix for loop positions

Site positions	8	9	10	11	12
A	2	76	2	10	3
C	83	6	1	10	4
G	8	3	91	8	86
U	2	10	1	67	2
Consensus	**C**	**A**	**G**	**U**	**G**

Acknowledgments

Work was supported in part by the RFBR (03-07-90181-в, 03-04-48469-a), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology", Russian Ministry of Industry, Sciences and Technologies (43.073.1.1.1501), SB RAS (Integration Projects No. 119), NIH USA (No. 2 R01-HG-01539-04A2), Russian Federal Research Development Program (contract No. 38/2004).

References

Addess K.J., Basilion J.P., Klausner R.D., Rouault T.A., Pardi A. Structure and dynamics of the iron responsive element RNA: implications for binding of the RNA by iron regulatory binding proteins // J. Mol. Biol. 1997. V. 274. P. 72–83.

Allerson C.R., Cazzola M., Rouault T.A. Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome // J. Biol. Chem. 1999. V. 274. P. 26439–26447.

Butt J., Kim H.Y., Basilion J.P. *et al*. Differences in the RNA binding sites of iron regulatory proteins and potential target diversity // Proc. Natl Acad. Sci USA. 1996. V. 93. P. 4345–4349.

Haile D.J. Regulation of genes of iron metabolism by the iron-response proteins // Am. J. Med. Sci. 1999. V. 318. P. 230–240.

Henderson B.R., Menotti E., Bonnard C., Kuhn L.C. Optimal sequence and structure of iron-responsive elements. Selection of RNA stem-loops with high affinity for iron regulatory factor // J. Biol. Chem. 1994. V. 269. P. 17481–1749.

Henderson B.R., Menotti E., Kuhn L.C. Iron regulatory proteins 1 and 2 bind distinct sets of RNA target sequences // J. Biol. Chem. 1996. V. 271. P. 4900–4908.

Rogers J.T., Randall J.D., Cahill C.M. *et al.* An iron-responsive element type II in the 5'-untranslated region of the Alzheimer's amyloid precursor protein transcript // J. Biol. Chem. 2002, V. 277. P. 45518–45528.

THE NEW APPROACH OF BOTH NEW AND OLD SEGMENTAL DUPLICATIONS SEARCH: REPETITIVE DNA AS A MOLECULAR ARCHAEOLOGY TOOL

Oparina N.*, Rychkov A., Mashkova T.

Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
* Corresponding author: e-mail: nixie@eimb.relarn.ru

Keywords: *alpha-satellite DNA, segmental duplications, RepeatMasker, recombination*

Summary

Motivation: The mechanism of segmental duplications remains unclear. The main approach for duplicon hunting is the pairwise alignment of "unique" (masked from the known repetitive elements) DNAs. The preferred centromere-proximal location of segmental duplications in the human genome implies the possible participation of (peri)centric repeats in their distribution. The high frequency of genomic repeats at the duplicons termini let us think, that there could be much more repeat-linked duplications.

*Results:*We used our own consensi for human satellite subfamilies for new human repetitive elements library (tested for use with the RepeatMasker program). Using modified Smith-Waterman method for detecting repetitive elements, we have produced the detailed repeats map in 34 build of the human genome. We have suggested and tested the new approach for detecting both recent and old duplicons, according to repeat maps pairwise comparison.

Availability: The RepeatsCompare program, human repeats library and test data are available upon e-mail request: nixie@eimb.relarn.ru.

Introduction

The pericentric organization of the human chromosomes is largely unknown. The central centromeric DNA consists of homogenized alpha-satellite arrays. Pericentric regions may contain a complex mixture of alpha-satellites, classical satellites, beta-, gamma-, 48-bp satellites (CER), chAB4 and other repeats, as well as great number of transposable elements. Suchregions of many chromosomes are enriched with segmental duplication events. Many additional segmental duplications will be found as a consequence of further human genome sequencing.Different tandem repeats are found at the termini or within the duplicons. Earlier we have found duplicated alpha-satellites using the full sequenced human 21cen-p genomic cosmid (Ac No AF105153). Duplicated satellites show higher frequency of structural rearrangements in comparison to "unique" DNA, complicating the duplicon termini analysis.

Methods and Algorithms

We have used the cross-match program for pairwise comparisons. The RepBase database of known human repeats was used for library. Unfortunately, there were only few consensi for alpha-satellite DNAs. We have selected different out-of-centromere alpha-satellite arrays (OOCA), produced monomer-by-monomer multiple alignments of all found alpha-satellites and created the additional set of consensi for diverged OOCA detection. The modified RepeatMasker program was used for mapping the repetitive elements on human chromosomes. The following procedure is used: 1) we select the parameters set, identical in case of paralogous DNAs comparison (repeat class/family (E.G., LINE/L1); repeat name (E.G., L1Hs); repeat orientation, positions in repeat consensus and the indels % between found repeats and their consensi); 2) we don't take in

consideration the substitutions between the found repeats and the consensi; 3) the minimal number of the potentially paralogous repeats in the segment is selected; 4) the distance variation between the successive potentially paralogous repeats in both genomic DNAs should not exceed 100 bp.

Implementation and Results

We have studied the OOCA arrays in the human genome. We have selected the set of human gap-free Genbank sequences exceeding 30 kb, containing more than 10 alpha-satellite monomers juncted to non-satellite DNA. After monomer-by-monomer multiple alignment the phylogenetic analysis of each alpha-satellite domain was performed. Revealed distinct clusters were used for consensi generation. At the second step different consensi were compared, the three types of them were detected and three supraconsensi for OOCA detection were computed. The library of 11 centromeric and 3 OOCA alpha-satellite consensi was created and adapted for using as a RepeatMasker library. At the Fig. 1 the phylogenetic tree is of different suprachromosomal centromeric alphoids (CnsW1-W5, CnsD1, CnsD2, CnsJ1, CnsJ2; dotted lines) form distinct cluster. The three OOCA clusters (1–3) are revealed and used for creating OOCA consensi.

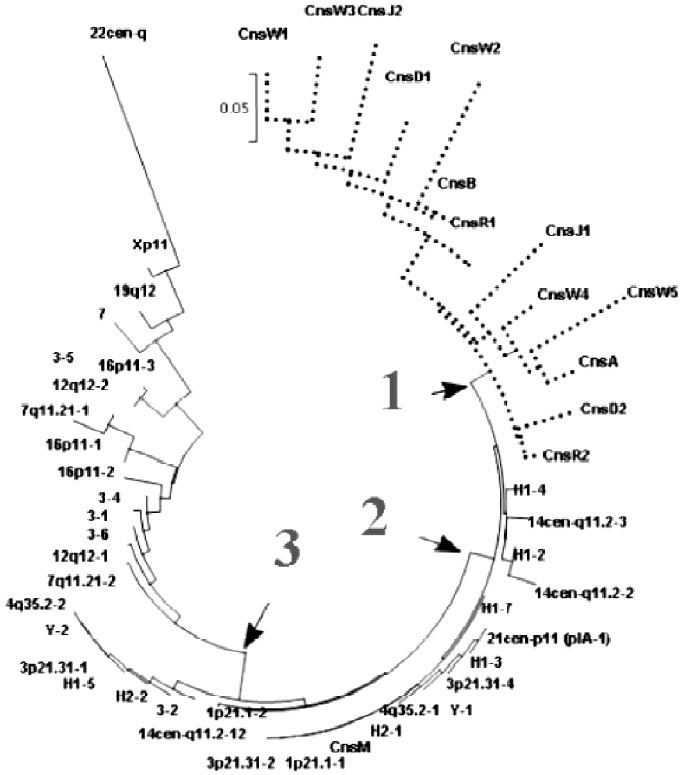


Fig. 1. The phylogenetic tree (NJ-tree created by MEGA software (boostrap value=1000), tested by PUZZLE) of centromeric and out-of-centromeric alpha satellite consensi.

The usual OOCA alphoid domain contains different alpa-satellite subfamilies, called subdomains. Most of such subdomains differs in their consensi monomers and form intermingled regions, usually containing insetions of other repeats. The example of OOCA domain structure is presented at the Fig. 2.

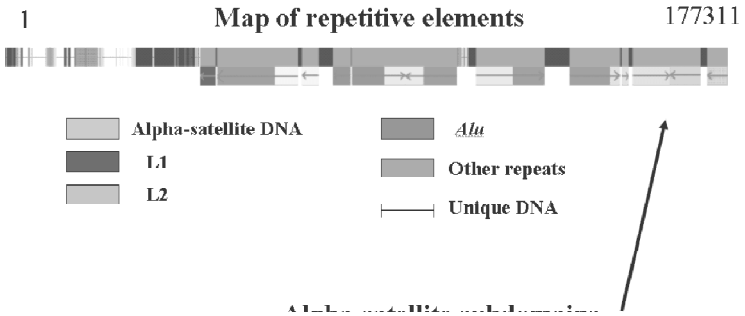


Fig. 2. The repetitive elements map and alpha-satellite subdomains distribution of the human OOCA segment (NCBI NT_025479 contig). Different aplha satellite subdomains are presented as greyscale. The inverted alpha satellite arrays are shown by arrows.

The created RepeatMasker library of alpha satellite consensi monomers were tested and could be used for mapping alpha-satellite families in human genomic DNA (see the fragment of our alpha satellite-specific RepeatMasker output for the NT_025479 at the Fig. 3).

```
74892 75058    267 356    89130 89296    1169 6.59 5.99 2.99     M(new) ALR 0 172
75060 75067    268 357    89298 89305    35 0 50 0               ALR-cons ALR 0 14
75061 75226    269 358    89299 89464    1046 8.43 7.83 4.22     R1 ALR 0 172
75227 75397    270 359    89465 89635    1214 6.43 4.09 3.51     M(new) ALR 0 172
75398 75568    271 360    89636 89806    1073 9.94 4.09 3.51     M(new) ALR 0 172
75569 75734    272 361    89807 89972    1110 7.23 6.02 2.41     M(new) ALR 0 172
75398 75568    271 325    84085 84255    1073 9.94 4.09 3.51     M(new) ALR 0 172
75569 75734    272 326    84256 84421    1110 7.23 6.02 2.41     M(new) ALR 0 172
75735 75903    273 327    84422 84590    1068 12.43 2.37 0.59    M(new) ALR 0 172
75904 76074    274 328    84591 84761    1121 7.02 4.68 4.09     M(new) ALR 0 172
76075 76246    275 329    84762 84933    1122 8.72 4.07 4.07     M(new) ALR 0 172
76247 76416    276 330    84934 85103    1085 8.82 5.88 4.71     R2 ALR 0 172
76588 76757    277 331    85275 85444    964 8.82 7.06 5.88      R2 ALR 0 172
76247 76416    276 365    90485 90654    1085 8.82 5.88 4.71     R2 ALR 0 172
76588 76757    277 366    90826 90995    964 8.82 7.06 5.88      R2 ALR 0 172
76758 76927    278 367    90996 91165    1157 10 1.18 0.59       M(new) ALR 0 172
76929 77098    279 368    91167 91336    1255 4.71 4.12 2.94     M(new) ALR 0 172
77099 77268    280 369    91337 91506    1116 10.59 2.35 1.18    M(new) ALR 0 172
```

Fig. 3. The mosaic structure of the alpha-satellite array. The RepeatMasker format of repeats map is given. The subdomains with different alpha satellite monomers consensi are shadowed.

Such subdomains in satellite DNA let us propose the involvement of recombination in the alpha-satellite mosaicism origin. Are any of these subdomains duplicated anywhere? We use the described approach for comparing the alpha-satellite-based repeat maps of the analysed Genbank sequences set. In 87 % cases we found the paralogous regions larger than 7 monomers. The dot matrix comparisons were carried out for paralogous sequences. Only 36 % cases contained recently (>90 % identity per 2 kb) duplicated alphoids. Most of found paralogous segments were not-distinguishable by homology-based pairwise comparison. For each pair we have produced the set of pseudoparalogues with the BLAST score same as for RepeatsCompare found paralogues. In 97 % these regions were more than 20 % differences in alpha-satellite monomer types and positions. Thus, we can reveal not only recent alpha-satellite duplicons, but the old ones too. The analysed set of OOCA DNAs contained a large amount of non-satellite regions. These loci are enriched with different types of known human repeats, including dispersed elements. The insertion events are well-known tool for phylogenetic trees reconstruction. We have tested the

RepeatsCompare approach for non-satellite DNAs, using the Repbase-based library of human repeats. We have selected the set of 20 pairs of sequences, contained recent alpha-satellite duplicons. In all cases we have found both old and recent (detected also by pairwise BLAST) paralogous segments.

Discussion

The segmental duplications and their involvement into alpha-satellite DNA mosaicism were described earlier (Oparina, Lakrua, 2003). The exclusion of (peri)centric chromosomal loci from genomic projects may lead to misunderstanding of the real involvement of (peri)centric repeats into segmental duplications. The modern approaches for duplicon hunting are not suitable for old or repeat-containing paralogues search (Lakrua, Oparina, 2003). Only the fraction of recent segmental duplications is taken into account: the arbitrary selected threshold of >90 % sequence identity (per 1–1,5 kb of "unique" DNA) is used for paralogues search. The event of repetitive element insertion is of high informability for phylogenetic analysis. The suggested approach could be informative for reconstruction of large-scale evolutionary history of duplications and recombinations in higher eukaryotes genomes.

References

Lakrua M.E., Oparina N.Iu., Mashkova T.D. Segment duplications in the human genome // Mol. Biol. (Mosk.). 2003. V. 37. P. 212–220.

Oparina N.Iu., Lakrua M.E., Rychkov A.A., Mashkova T.D. Tandem and interspersed repeats contribute to the mosaic structure of segmented duplications in the human genome // Mol. Biol. (Mosk.). 2003. V. 37. P. 228–233.

NUCLEOSOME POSITIONING SIGNAL ANALYSIS

Orlov Yu.L.*, Levitsky V.G.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: {orlov,levitsky}@bionet.nsc.ru

Keywords: *nucleosome, gene expression regulation, context signals*

Summary

Motivation: Computer analysis of context features in nucleosome formation sites is of a great importance. The aim was to examine the variable memory Markov model method for estimation of nucleosome formation potential (tendency to bind histone octamer and nucleosome forming) of arbitrary nucleotide sequences.

Results: The variable memory Markov models method was applied for estimation of the nucleosome formation ability and showed quite similar results on genomic DNA as previously applied method of discriminant analysis of oligonucleotide frequencies RECON. An analysis of phased nucleotide sequences containing nucleosome formation sites revealed periodic signals in local text complexity.

Availability: Software is available by request to the corresponding author.

Introduction

Chromatin from various genomic regions is, as a rule, represented by regular arrays of nucleosomes (Aalfs, Kingston, 2000; Becker, 2002). The neighboring nucleosomes are connected by linker DNA with a length ranging from 20 to 80 bp. The sequence-directed nucleosome positioning plays an important functional role in providing a proper interaction of DNA functional sites with non-histone proteins. The mechanisms of sequence-directed nucleosome positioning have been studied in numerous experiments both *in vivo* and *in vitro*; the results obtained suggest the existence of a specialized chromatin (nucleosome) code determining such positioning through multiple histone-DNA interactions (Trifonov, 1997).

Various research teams have succeeded in discovering a number of periodic contextual and conformational signals and rules regulating nucleosome positioning (Trifonov, 1997; Kiyama, Trifonov, 2002; Levitsky *et al*., 2004). Although this field is intensely studied, the mechanisms underlying nucleosome positioning are yet far from being clearly understood. Computer analysis of nucleosome positioning code needs different unrelated methods, such as estimation of sequence linguistic complexity, search for periodic signals and Markov models. Indeed, context nucleosome positioning code (Trifonov, 1997) assumes degeneracy (quite different DNA sequences are able to interact with histone octamer and nucleosome formation), weakness of context signals, and absence of clear signal localization. Markov model (Orlov *et al*., 2002) of text generation is based on preceding symbols, thus as a statistical model Markov model corresponds to theoretical assumptions about nucleosome code.

New prediction method is based on Variable memory Markov model and allows detect the prediction tendency to bind histone octamer for arbitrary nucleotide sequences. The software developed allows to reveal context features such as periodicity and low complexity region distribution in DNA sequences containing nucleosome formation sites.

It was compared with the software developed previously, which allows predicting nucleosome formation potential, by dinucleotide frequencies (Levitsky *et al*., 2001; Levitsky *et al*., 2004).

Methods

To develop programs for detecting and recognizing nucleosomal context and to evaluate their prognostic capability, several samples of NFS – DNA sequences with experimentally demonstrated ability to form nucleosomes – are used. They include 141 sequences extracted from the GenBank/ EMBL databank according to the accession numbers and positions indicated in the database Nucleosomal DNA (Ioshikhes and Trifonov, 1993). These samples were used for designing software packages (Levitsky *et al*., 2001), intended for analysis of contextual NFS properties and prediction of the ability of arbitrary DNA sequences to form nucleosomes.

Analysis of text complexity of NFS was done using "LowComplexity" software for estimation of DNA sequence complexity by several methods including modified Lempel-Ziv method and entropy estimations (Orlov *et al*., 2004).

Prediction of nucleosome formation potential was developed using extended Markov model trained on the database. We consider Variable Memory Markov Models for the generation of symbols based on a stationary source. The local preceding context (1–9 bp) defines the current state of the Markov model independent of the position in the text. Such model is based on suffix tree. Suffix trees as a basis for text generation were studied for proteins as an alternative to the HMM in the form of a sub-class of probabilistic finite automata.

Results and Discussion

Prediction of nucleosome formation potential. Prediction of nucleosome formation potential was preformed by estimation of probability of arbitrary sequence correspond to the Markov model trained on the database. Fit function was constructed as minus logarithm of probability to obtain a sequence S in sliding window by fixed length n: -Prob(S_n). Testing the software on control sample of experimentally defined NFS showed correct results. Nucleosome formation is especially important in regulatory regions of genes, in particular, promoters (Aalfs, Kingston, 2000). We analyzed set of gene promoters from TRRD database (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd). Averaged profiles are presented at Figure 1.

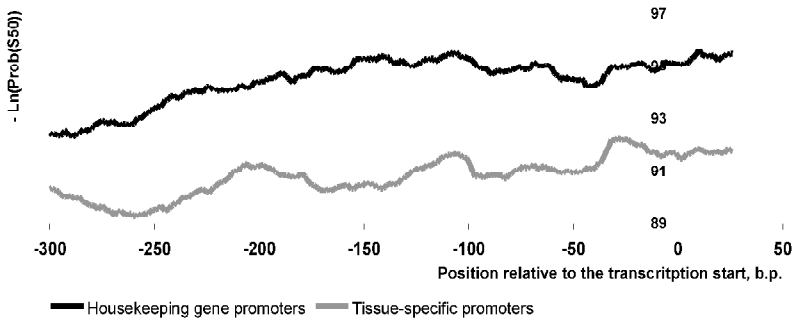


Fig. 1. Nucleosome formation potential profiles in sliding window 50 bp for sets of eukaryotic gene promoters phased relative to the transcription start [-300;+50]. Lower level means greater similarity to nucleosome formation sites. Averaged profile for the set of promoters of housekeeping genes (high expression level) and tissue-specific genes (lower expression level) are designated by solid black and light gray lines correspondingly.

Comparison of nucleosome potential function based on Variable Memory Markov Models and function based on dinucleotide frequencies and discriminant analysis (Levitsky *et al*., 2001) revealed the similar results for nucleosome formation site sequences and promoter sequences.

Promoters of housekeeping genes have less preference to contain NFS than tissue-specific genes (Levitsky *et al*., 2001). Thus the research suggested that high expressed genes in nucleus should be free of chromatin packing than rare expressed (tissue-specific) genes.

The result is in good accordance with our previous analysis (Levitsky *et al*., 2001). Moreover significant correlation of nucleosome potential estimation functions was shown on genomic DNA. In addition our analysis shown than introns and 5'-non-translated gene regions have less nucleosome formation potential than exons and regulatory regions of gene expression.

Low complexity regions in nucleosome formation sites. Let us consider local complexity profiles by sliding window for phased set of NFS. Profile value for every sequence reflects number of copying operation (direct repeats) to construct the sequence. It is an integer value; we averaged all this values for every position of the phased sequence set. The DNA sequences by length 400 bp were phased relative to the center of NFS [-200;+200].

Middle part [-73;+73] corresponds to 147-bp DNA wrapped around histone octamer (gray bar in Fig. 2). Linker DNA sequences (approximately 50 bp) are designated by stripped bars to left and to right from the center. Since we assume existence of ordered nucleosome array in genomic DNA we designate far left and right flanks of sequence by gray bars as part of neighboring nucleosome sites.

Figure 2 presents only averaged complexity profile by modified Lempel-Ziv method in sliding window 27 bp for whole sequence set. One can see, that complexity values are minimal at the flanks of core nucleosome site (site position designated by gray bar). Linker DNA corresponds to higher level of text complexity. Moreover, local trend of increase and decrease of complexity from NFS center exist (dotted line). At whole we see symmetrical picture of complexity values distribution (Fig. 2) corresponding to biophysical assumptions about molecular mechanisms of DNA packing. In addition, profile complexity shows periodical distribution of local complexity minima (Fig. 2, vertical arrows). These minima have period 10–11 bp corresponding to previous estimations (Kiyama, Trifonov, 2002; Ioshikhes, Trifonov, 1993). Such preference in local minima distribution could be connected with periodic distribution of simple sequence repeats, even as short as 2–3 nucleotides.

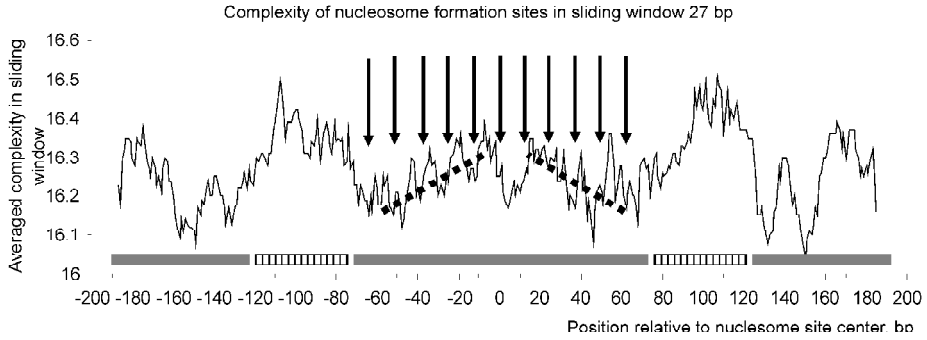


Fig. 2. Averaged complexity profile by modified Lemple-Ziv method for phased set of nucleosome formation sites [-200;+200]. Gray bars indicate 146 bp core histone binding sequences, stripped bars correspond to linker DNA. Profile trends are indicated by straight dotted lines. Arrows show periodic (10–11 bp) local complexity minima.

We plan further developing of an integrated information system Nucleosomal DNA Organization, comprising Nucleosome Positioning Region Database and software packages for nucleosome formation sites (NFS) recognition (Levitsky *et al*., 2001).

Acknowledgements

The authors are grateful to Yu.D. Oshchepkov, A.V. Katokhin, O.A. Podkolodnaya and N.A. Kolchanov for valuable discussion. This work was supported in part by the RFBR (02-07-90355, 03-04-48555), Ministry of Education (E 02-6.0-250), SB RAS (Integration Projects 119), NATO (LST.CLG 979815), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology".

References

Aalfs J.D., Kingston R.E. What does 'chromatin remodeling' mean? // Trends Biochem Sci. 2000. V. 25. P. 548–55.

Becker P.B. Nucleosome sliding: facts and fiction // EMBO J. 2002. V. 21. P. 4749–53.

Ioshikhes I., Trifonov E.N. Nucleosomal DNA sequence database // Nucleic Acids Res. 1993. V. 21. P. 4857–9.

Kiyama R., Trifonov E.N. What positions nucleosomes? – A model // FEBS Lett. 2002. V. 523. P. 7–11.

Levitsky V.G., Katokhin A.V., Podkolodnaya O.A., Furman D.P. Nucleosomal DNA organization: an integrated information system // Bioinformatics of Genome Regulation and Structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London, 2004. P. 3–12.

Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis // Bioinformatics. 2001. V. 17. P. 998–1010.

Orlov Yu.L., Potapov V.N., Filippov V.P. Recognizing functional DNA sites and segmenting genomes using the program Complexity // Proc. of the BGRS'2002. Novosibirsk: IC&G, 2002. V. 3. P. 243-246.

Orlov Yu.L., Potapov V.N. Complexity: Internet-resource for analysis of DNA sequence complexity // Nucleic Acids Res. 2004. (web-server issue 2004). In press.

Trifonov E.N. Genetic level of DNA sequences is determined by superposition of many codes // Mol. Biol. (Mosk). 1997. V. 31(4). P. 759-67.

COMPUTER ANALYSIS OF GENOMIC SEQUENCE COMPLEXITY: NEW APPLICATIONS

***Orlov Yu.L.*[1]*, *Potapov V.N.*[2], *Poplavsky A.S.*[1]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia

* Corresponding author: e-mail: {orlov,apost}@bionet.nsc.ru; vpotapov@math.nsc.ru

Keywords: *computer analysis, complexity*

Summary

Motivation: The search for DNA regions with low complexity is one of pivotal tasks of modern structural analysis of complete genomes. The low complexity may be preconditioned by strong inequality in nucleotide content (biased composition), by tandem or dispersed repeats, by palindrome-hairpin structures, as well as by combination of all these factors.

Results: Several numerical measures of text complexity including combinatorial and linguistic ones, together with complexity estimation by modified Lempel-Ziv algorithm, were implemented in a software tool "LowComplexity". The software developed enables to search for low complexity regions in long sequences, i.e., complete bacterial genomes or eukaryotic chromosomes. In addition, it estimates complexity of groups of aligned sequences. By comparison of complexities of various functional regions, we have demonstrated that complexity of introns and regulatory regions is less than in coding regions.

Availability: http://wwwmgs.bionet.nsc.ru/programs/low_complexity

Introduction

Analysis of genomic sequences issues the challenge to search for the regions with the low text complexity, which could be functionally important (Hancock, 2002; Wan *et al*., 2003). Low complexity regions are often treated as the regions of biased composition containing simple sequence repeats. The sequence enriched with imperfect direct and inverted repeats may be also considered as the sequence with low complexity.

Complexity of symbolic sequence reflects an ability to represent (compress) a sequence in a compact form based on some structural features of this sequence. In the software suggested, we have realized several additional estimates of complexity in order to compare different approaches. In particular, following evaluations of the text complexity are presented: 1) by nucleotide frequency content (Wootton, Federhen, 1996); 2) by entropy of the given order of words (oligonucleotides); 3) by linguistic complexity (Troyanskaya *et al*., 2002); 4) modifications of complexity measure by Lempel and Ziv (Gusev *et al*., 1999). As the method for complexity evaluation, we have chosen the scheme of the text representation in terms of repeats, which uses the concept of complexity of a finite symbolic sequence, introduced by Lempel and Ziv (Lempel, Ziv, 1976). The Lempel–Ziv complexity measure is based on text segmentation; we termed it as complexity decomposition (Gusev *et al*., 1999). It may be interpreted as representation of a text in terms of repeats. Based on this approach, the Internet-available tools LZcomposer was presented (Orlov *et al*., 2004; (http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/).

The software is designed for effective search for the regions with low complexity in extended DNA sequences. The search is provided by different methods and its operation time is linearly dependent upon the sequence length by application of *l*-gram trees for the sequence representation. Modified Lempel-Ziv method for complexity estimation was shown as most sensible for low complexity region detection.

The software enables to calculate average complexity profile for the sets of sequences in the FASTA format. By using the system, we have demonstrated the higher nucleotide sequence complexity of exons and lower complexity of introns (Orlov *et al*., 2004). Also, we have found the alteration of the local text complexity for the splicing sites.

Methods

Complexity estimation by Lempel and Ziv scheme. Lempel and Ziv proposed to measure complexity of sequence by the number of steps of generating process (Lempel, Ziv, 1976). The permitted operations here are generation of a new symbol (this operation is necessary at least to synthesize the alphabet symbols) and direct copying of a fragment from the already generated part of the text. Copying implies search for a prototype (repeat in a common sense) in the text and extension of the text by attaching the 'prepared' block. We use direct and inverted repeats as the prototypes. A user may choose any of the types of copying operation.

The scheme for generating the sequence S may be represented as a concatenation H of the fragments:

$$S = S\,[1: i1]\; S\,[i1+1: i2] \ldots S\,[ik - 1+1: ik] \ldots S\,[im - 1+1:N]$$

$$H(S) = S\,[1 : i1],\; S\,[i1+1 : i2], \ldots S\,[ik - 1+1 : ik], \ldots S\,[im - 1+1 : N], \qquad (1)$$

where S [ik – 1+1 : ik] is the fragment (component) generated at the k-th step (a sequence of elements located from the position ik – 1 + 1 to ik); N, the length of sequence; and $m=m_H(S)$, the number of steps generating the process. The scheme with minimal number of steps m should be selected. This scheme determines the complexity of sequence S:

$$CLZ(S) = \min_H \{m_H(S)\}. \qquad (2)$$

The algorithm implementation for DNA research was described in details in (Gusev *et al*., 1999).

We construct the complexity profile in the sliding window with the length N, the evaluation of complexity is calculated as the whole number *CLZ*(S) of components of complexity decomposition in the window N, or as the relative number of the components *CLZ*(S)/N.

In the software presented, we have realized also the algorithm for evaluation of the word complexity in accordance with the nucleotide frequency (Wootton, Federhen, 1996). The algorithm is used as BLAST search preprocessing for masking the low complexity regions. Evaluation of complexity in a text region *CWF* by Wootton and Federhen (Wootton, Federhen, 1996) is realized by the formula

$$CWF = (1/N)\, log_K\, (N! \,/\, \prod_{i=1}^{K} n_i!), \qquad (4)$$

where N is the window size, n_i is the number of nucleotide in a window, K is the alphabet size (for DNA, K=4).

Another realized method for estimating complexity is evaluation of the entropy *CE*. In the same designation as in (4) entropy is presented as

$$CE = -\sum_{i=1}^{N} (n_i / N)\, log_K (n_i / N). \qquad (5)$$

As the logarithm is taken by the basis K, the complexity values fall to the interval [0;1].

The complexity may be estimated by the entropy measures, including entropy in the Markov model of the high order given below:

$$CM = - \sum_{i=1}^{M} (m_i/(N-m+1)) \log_M (m_i/(N-m+1)), \quad (6)$$

where m_i are the number of oligonucleotide in a window, $M=K^m$ is the total number of words (oligonucleotides) with the length m, K is the alphabet size.

The linguistic complexity accounts for the occurrence of long words (11). Quantitatively, the linguistic complexity, *CL*, is determined as the ratio of the sum of sub-words (sub-lines) number occurring in the sequence analyzed to the maximal number of all these sub-words. Thus, the number of different oligonucleotides (sub-words) with the length i, $1<i\leq m$ that were found in the window equals to

$$CL = (\sum_{i=1}^{m} V_i) / (\sum_{i=1}^{m} V_{max\,i}). \quad (7)$$

Here V_i is the number of oligonucleotides with the length i, K is the alphabet size, m is the maximal length of a sub-word (oligonucleotide) $1\leq m\leq N$. $V_{max\,i}$ is the maximal possible number of oligonucleotides with the length i, for the window with the ordered size N, $V_{max\,i} = \min(K^i, N-i+1)$. For example, in the window with the size of 20 bp, it is possible to displace all of 4 nucleotides, all 16 dinucleotides, 18 trinucleotides (out of 64 trinucleotides, only 18 may be input into the sequence with the length 20), etc., including 2 sub-words with the length 19 and one sub-word with the length 20 bp.

Thus, we achieve $\sum_{i=1}^{20} V_{max\,i} = 4+16+18+17+16+15+...+2+1=191$.

CL varies in the interval [0;1]. In order to limit the usage of long sub-words, the summation may be limited by the parameter m, $m\leq N$. The linguistic complexity gives evidence about variability of sub-words, but it is not suitable for searching for particular repeats and their localization.

Let us consider an example of calculation of linguistic complexity for the nucleotide sequence containing the AP2 transcription factor binding site, *GTGCCCCGCGGGAACCCCGC* with the length N=20:

Sub-word length	1	2	3	4	5	6	7	...	19	20
Number of possible sub-words	4	16	18	17	16	15	14	...	2	1
Number of sub-words found	4	9	13	14	14	14	14	...	2	1

Totally, 173 sub-words were found, 191 sub-words could be possibly found. By formula (7), linguistic complexity *CL* equals to 0.906(=173/191).

For the example sequence, all complexity evaluations are given below:

Complexity estimation	CLZ	CLZ/*N*	CE	CM	CWF	CL
Complexity value	10	0.5	0.789	0.706	0.650	0.906

Results and Discussion

The program software developed is designed to analyze the groups of sequences, or to calculate the mean value of complexity profile at each position. This analysis evaluates alteration of text complexity for the functional groups. A user may choose the mode of program implementation: (i) for analysis of a single extended sequence by different methods or (ii) analysis of a group of relatively short, up to 1Kb in length, sequences. A table of complexity values is constructed for a

window, of ordered size N, sliding along the sequence. The sequence complexity is assigned to a window center. To illustrate the mode how to operate with a group of phased sequences, we have analyzed the set of acceptor and donor splice sites extracted from the database SpliceDB (Burset *et al*., 2001). We find regularities that are common for the splicing sites (Fig.).

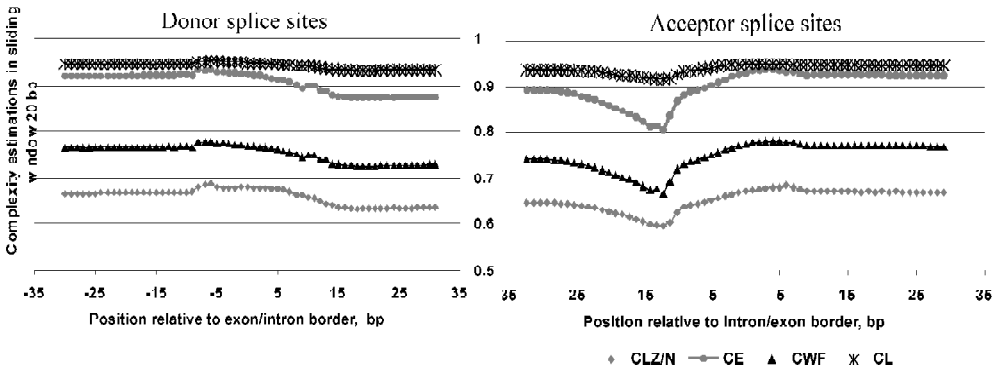


Fig. The averaged complexity profiles in the sliding window 20 bp for the sets of donor (a) and acceptor (b) splicing sites in mammals. The profile of linguistic complexity (*CL*), entropy (*CE*), complexity estimation by Wootton-Federhen (*CWF*), Lempel-Ziv complexity (*CLZ*) are shown. By abscissa, position of the window center is shown relatively the border between exon and intron (for donor sites) and from intron to exon (for acceptor sites). The sequence length equals to 82 bp.

The sequences were of 82 nucleotides in length with canonical dinucleotides GT and AG marking the border between exon and intron in the center. We have calculated complexity within the sliding window of 20 bp. The profile was constructed by averaging the complexity values along all sequences of the set in a window. The step of sliding window equals to one position. Mean values of profiles for the sets of donor and acceptor splicing sites are overlaid. By abscissa, the window positioning relatively canonical dinucleotide is given. As seen, the average complexity in the sliding window for acceptor splicing sites increases in direction from intron to exon. On the contrary, for the donor splicing sites, the complexity drops down. Hence, the coding regions are characterized by higher complexity than non-coding regions. The decrease in complexity for acceptor splicing sites within the region [-20;-10] relatively the border between exon and intron points out that the structure is conserved.

Note that for revealing structures of extended imperfect repeats, the Lempel-Ziv complexity and linguistic complexity estimates are the most suitable. The entropy and complexity values by Wootton and Federhen are strongly correlating, with the correlation coefficient equaling to 0.95–0.99 for genome sequences analyzed. However, Lempel-Ziv complexity values and linguistic complexity values are less correlated with entropy estimates.

The results of complexity analysis of the phased sites of splicing, exons, introns, and promoters are Internet-available via the address http://wwwmgs.bionet.nsc.ru/mgs/programs/lzcomposer/ResPromoters.htm.

Comparison of several methods evaluating text complexity is presented for the first time.

By comparison of complexities of various functional regions, we have demonstrated that complexity of introns and regulatory regions is less than in coding regions. Our results support previously obtained data (Troyanskaya *et al*., 2002) for bacterial genomes that linguistic complexity differs between the coding and non-coding gene regions. We have proved that this regulation is also valid in eukaryotes by studying gene region complexity by several other complexity measures.

Acknowledgements

This work was supported in part by the RFBR (01-07-90376, 02-07-90355, 03-04-48506), Russian Ministry of Education (E 02-6.0-250), NATO (LST.CLG 979815), SB RAS (Integration project No. 119), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology".

References

Burset M., Seledtsov I.A., Solovyev V.V. SpliceDB: database of canonical and non-canonical mammalian splice sites // Nucleic Acids Res. 2001. V. 29. P. 255–259.

Gusev V.D., Nemytikova L.A., Chuzhanova N.A. On the complexity measures of genetic sequences // Bioinformatics. 1999. V. 15. P. 994–999.

Hancock J.M. Genome size and the accumulation of simple sequence repeats: implications of new data from genome sequencing projects // Genetica. 2002. V. 115. P. 93–103.

Lempel A., Ziv J. On the complexity of finite sequences // IEEE Trans. Inf. Theory. 1976. IT-22. P. 75–81.

Orlov Yu.L., Filippov V.P., Potapov V.N., Kolchanov N.A. Construction of stochastic context trees for genetic texts // In Silico Biology. 2002. V. 2. P. 233–247, <http:www.bioinfo.de/isb/2002/02/0022/>.

Orlov Yu.L., Gusev V.D., Miroshnichenko L.A. LZcomposer: decomposition of genomic sequences by repeat fragments // Biophysics (Mosk.). 2004, (in Press).

Troyanskaya O.G., Arbell O., Koren Y., Landau G.M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity // Bioinformatics. 2002. V. 18. P. 679–688.

Wan H., Li L., Federhen S., Wootton J.C. Discovering simple regions in biological sequences associated with scoring schemes // J. Comput. Biol. 2003. V. 10. P. 171–185.

Wootton J.C., Federhen S. Analysis of compositionally biased regions in sequence databases // Methods Enzymol. 1996. V. 266. P. 554–71.

CONTEXT FEATURES OF TRANSCRIPTION FACTOR BINDING SITE SEQUENCES: RELATION TO DNA-BINDING DOMAIN CLASSIFICATION

***Orlov Yu.L.*[1]*, *Proscura A.L.*[1], *Vityaev E.E.*[2], *Arrigo P.*[3]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia, e-mail: vityaev@math.nsc.ru; [3] ISMAC, via De Marini 6 16149 Genova, Italy, e-mail: arrigo@ge.ismac.cnr.it

* Corresponding author: e-mail: {orlov,anya}@bionet.nsc.ru

Keywords: *transcription factor binding sites, DNA-binding domains, classification, gene expression regulation, context features, complexity*

Summary

Motivation: Classification of eukaryotic transcription factor binding sites (TFBS) by context features of DNA sequences is of importance for analysis gene transcription regulation. Growth of information volume for gene regulatory sequences makes it possible to reveal new statistical regularities governing DNA-binding and gene expression regulation.

Results: We search for sequence constraints connected with text complexity for core regions of transcription factor binding sites. The content of protein-binding nucleotide sequences is connected with DNA-binding domain classification. These finding suggest new approaches for TFBS classification and clusterization.

Availability: the software available at http://wwwmgs.bionet.nsc.ru/mgs/programs/low_complexity, database at http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/, results and supplementary materials are available by request to the authors.

Introduction

The binding sites of sequence specific transcription factors are an important and relatively well-understood class of functional non-coding DNA. Although a wide variety of experimental and computational methods have been developed to characterize transcription factor binding sites, they remain difficult to identify (Moses *et al*., 2003).

Let us consider problem of context differences in structure of transcription factor binding site (TFBS) core and flanking regions. We used TFBS of the length 100 bp from TRRD database. 67 different sets of TFBS were prepared (Kolchanov *et al*., 2002). Each sample contains 15–100 sequences. We consider phased set of sequences, i.e. by equal length without gaps.

Since these DNA sites possess common property to bind protein, they have similar structural restrictions on nucleotide content, flexibility and other physical-chemical parameters of DNA helix. Hence, the sequences should be similar. Similarity of sequences allows construct multiple alignment. But since core protein-binding sequences are relatively short, 4–10 bp, we cannot expect high similarity level.

The experimental data make evident an important role of DNA conformational features for site functioning. Previous research revealed conservation of some physical-chemical dinucleotide properties for TFBS sequences (Ponomarenko *et al*., 1999).

We implement several methods for assessing complexity of nucleotide sequence (Orlov, Potapov, 2004) to reveal low and high complexity regions in TFBS.

The task is to find common context features differing for core region and flanking regions.

Materials and Methods

In order to find common features of nucleotide sequences containing TFBS we construct complexity profile in sliding window. We construct profile in sliding window for every sequence separately. Sliding window should be enough short and comparable to core protein-binding region. We used 15–20 bp window sizes.

Sequences are phased relative to the center of core region. Then we calculate for every position of phased sequence average, minimal and maximal values of complexity profile of sequences in the set.

Multiple alignment of TFBS with shorter flanking regions show, in average, higher level of conservation than alignment of TFBS with longer flanking regions (Results are not shown).

We used several methods for text complexity estimation: entropy, entropy of words, complexity by Lempel-Ziv method and linguistic complexity (Orlov, Potapov, 2004, http://wwwmgs.bionet.nsc.ru/mgs/programs/low_complexity).

Implementation and Results

As we show, total TFBS could be subdivided into two large groups: 1) TFBS with higher text complexity of core region than flanking regions; 2) TFBS with lower text complexity of core region than flanking regions.

First class with higher text complexity contains majority of TFBS analyzed (more than 30 sets).

Second class, with lower complexity of core region, contains 18 TFBS.

The level of text complexity for flanking regions is the same for both classes. Main contribution in difference of complexity level is due to core regions. The difference is connected with protein transcription factor structure and regularity region evolution.

We construct complexity profile for representative set containing 67 different TFBS samples.

When we have analyzed averaged complexity profiles of all TFBD samples we have three distinct clusters. First cluster is characterized by higher average complexity of core region, second cluster – by lower average complexity, third cluster - by variable complexity of core region.

Figure shows averaged values for complexity profile of TFBS sequence samples with clear maximum (Fig.a) or minimum (Fig. b) in the site center.

Thus we see two different groups of TFBS – first group with higher complexity, and second group with lower complexity.

It was shown that for both group of TFBS namely mean values in core regions are different, while the values in flanking regions are approximately equal (see Fig.). Similar qualitative results we obtained by other measures like linguistic complexity (Troyanskaya *et al*., 2002) and entropy estimations. The estimations were done in [0;1] scale (Orlov, Potapov, 2004). All complexity estimations by different methods were correlated on the nucleotide sequences containing TFBS analyzed.

To characterize difference in complexity profiles let us consider classification of protein factors binding these DNA sequences (Wingender, 1997). TFBS with higher DNA sequence complexity mainly related to Superclass 1 (Basic domains). Proteins CEBP, CEBPalpha, CEBPbeta, CREB, CRE, JunD, JunFos, NFE2 related to first superclass (Basic domains), namely to class 1 (Leucin zipper, bZIP). Proteins CLOCK/BMAL1, cMyc are also related to Superclass 1 (class Helix-Turn-Helix, (bHLH-ZIP) 1.3). Protein ARNT is also related to this class (class 1.2, Helix-Turn-Helix, bHLH).

At the same time proteins DBP and NFkB related to 4 superclass (beta-Scaffold Factors with Minor Groove Contacts), (class 4.1, RHR – Rel homology region). Superclass 2 (Zinc-coordinating DNA-binding domains) contains protein ER and PPRE (2.1, Cys4 zinc finger of nuclear receptor type) and GATA1 (2.2) (diverse Cys4 zinc fingers). HSF1 related to Superclass 3 (Helix-turn-helix) and 3.4 class: Heat shock factors).

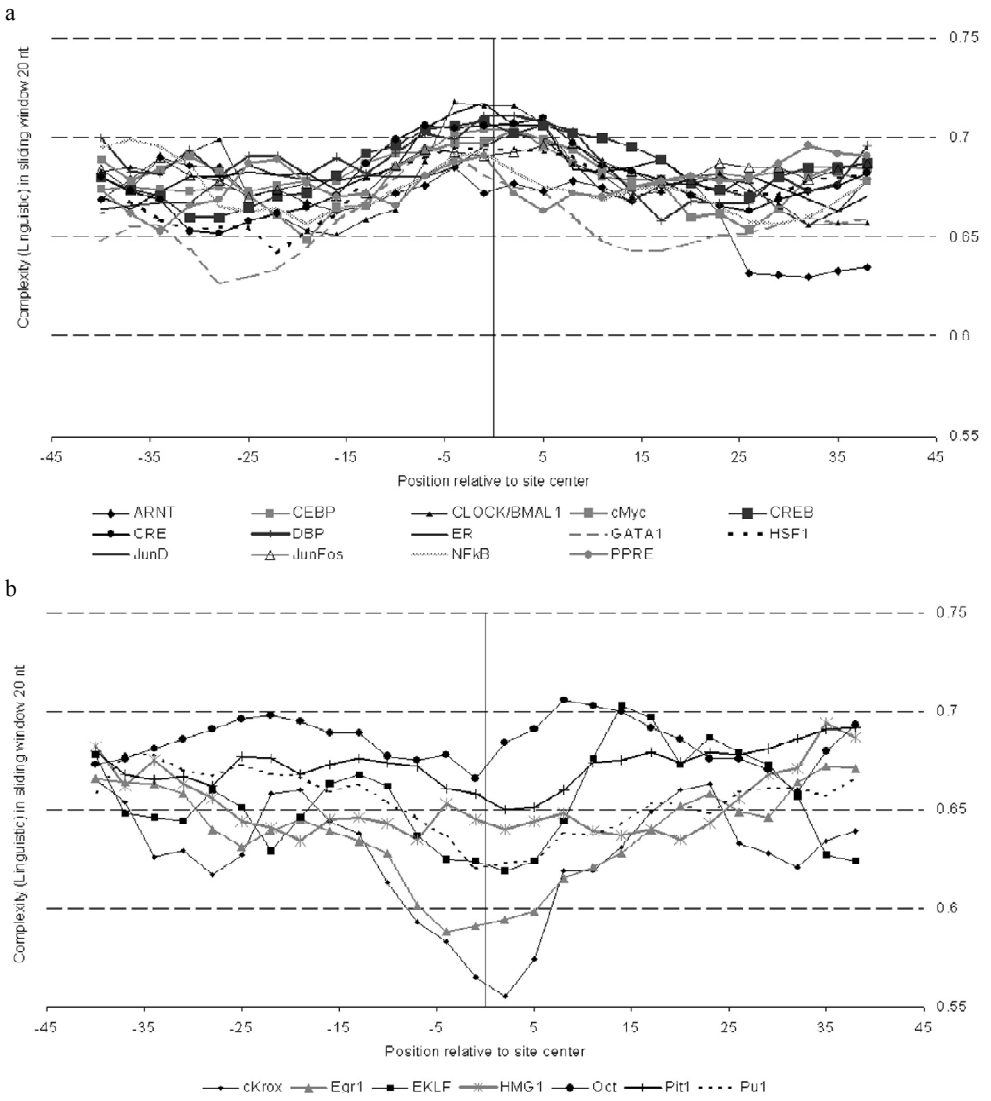


Fig. Averaged profiles of linguistic complexity in 20 bp window for phased TFBS sequences. We count all subwords up to length 5 nt. Sequences are phased [-50;+50] relative to the center of site annotated in TRRD database. Position corresponds to center of 15 bp window. Thick red line indicate averaged complexity profile for all samples (average by averages in samples).
(a) Profiles for TFBS with higher complexity on core region than in flanking sequences.
(b) Profiles for TFBS with lower complexity on core region than in flanking sequences.

TFBS with lower complexity of core region bind proteins not related to first cluster. These are: cKrox, Egr1, EKLF, Sp1 and Sp3 (2.3 Class: Cys2His2 zinc finger domain), HMG1(high-mobility-group protein 1; Class 0: not classified), Oct and Pit1 (3.1 Class: Homeo domain), Pu1 (3.5 Class: Tryptophan clusters).

Conclusion and discussion

Majority of TFBS samples analyzed contains sites with higher complexity of core region than flanking sequences. This set sequences also includes sites with several maximal values in 100 bp region, i.e. maximum in center and high values in flanking regions at 20–30 bp upstream or downstream. It hard to classify by DNA sequence complexity some TFBS like E2F. E2F is in 1.3 class (bHLH-ZIP) family 1.3.2 Family: Cell-cycle controlling factors). Complexity profile for this TFBS has not clear minimum.

Cluster of TFBS with lower complexity has only 8 sites with clear minimal averaged complexity in core region. We could add to the cluster also 10 sites with several regions of lower complexity.

It is interesting to note that TF with low complexity of corresponding binding sites (Oct, EKLF, HMG) capable to bind DNA within nucleosome or with higher predicted nucleosome formation potential (Levitsky, 2004). Complexity of TFBS sequence could be connected with nucleosome formation potential of promoters containing these sites.

The statistical constraints on complexity could be used as signals for TFBS recognition in genomic DNA.

Acknowledgements

Authors are grateful to V.G.Levitsky and N.A.Kolchanov for valuable discussion. This work was supported in part by the RFBR (02-07-90355, 03-07-96833, 03-04-48506), Russian Ministry of Industry, Science and Technology (43.073.1.1.1501), NATO (LST.CLG 979815), SB RAS (Integration Project No. 119), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”.

References

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription regulatory regions database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. V. 30(1). P. 312–7.

Levitsky V.G. RECON: a program for prediction of nucleosome formation potential // Nucleic Acids Res. 2004, (in press).

Moses A.M., Chiang D.Y., Kellis M., Lander E.S., Eisen M.B. Position specific variation in the rate of evolution in transcription factor binding sites // BMC Evol. Biol. 2003. V. 3(1). P. 19.

Orlov Yu.L., Potapov V.N. Complexity: Internet-resource for analysis of DNA sequence complexity // Nucleic Acids Res. 2004. (web-server issue 2004) In press.

Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Podkolodny N.L., Savinkova L.K., Kolchanov N.A., Overton G.C. Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins // Bioinformatics. 1999. V. 15(7/8). P. 687–703.

Troyanskaya O.G., Arbell O., Koren Y., Landau G.M., Bolshoy A. Sequence complexity profiles of prokaryotic genomic sequences: a fast algorithm for calculating linguistic complexity // Bioinformatics. 2002. V. 18. P. 679–688.

Wingender E. Classification of eukaryotic transcription factors // Mol. Biol. (Mosk). 1997. V. 31(4). P. 584–600. (In Russian).

SITECON: A TOOL FOR TRANSCRIPTION FACTOR BINDING SITE RECOGNITION

Oshchepkov D.Yu.*, Grigorovich D.A., Ignatieva E.V., Khlebodarova T.M.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: diman@bionet.nsc.ru

Keywords: *Transcription factor binding sites, conformational and physicochemical DNA properties, site recognition*

Summary

Motivation: The local DNA conformation in the region of transcription factor binding sites determined by context is one of the factors underlying the specificity of DNA–protein interactions. Analysis of local conformation of a set of functional DNA sequences may allow for determination of the conservative conformational and physicochemical parameters reflecting molecular mechanisms of interaction, this data can be effectively used for site recognition.

*Results:*The Web resource SITECON is designed to detect conservative conformational and physicochemical properties in transcription factor binding sites, contains the knowledge base of conservative properties for over 100 high quality samples of sites, and allows for recognition of potential transcription factor binding sites basing on these conservative properties from both the knowledge base and results of analysis of a sample proposed by an user.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/.

Introduction

An increasing volume of experimental data suggests that the function of transcription factor binding sites is determined to a considerable degree by the context-dependent conformational and physicochemical DNA properties (CDCPP) (Starr *et al.*, 1995; Meierhans *et al.*, 1997). An ever-increasing data of structural analyses demonstrate both heterogeneity of conformational and physicochemical properties and their dependence on the nucleotide sequence (Suzuki *et al.*, 1997). Moreover, it was been previously shown that CDCPP might be significant for the site function. The developed computer system ACTIVITY (Ponomarenko *et al.*, 1999) demonstrated a successful application of CDCPP for site activity prediction. That is why we believe that CDCPP may be an alternative approach to detection of specific features of transcription factor binding sites and their successful recognition.

Thus, the local conformation of DNA molecules determined by the context is a factor affecting the specificity of DNA-protein recognition. This suggests that certain conformational and physicochemical properties of the variants of genomic sequences interacting with a certain regulatory protein should be preserved. At the same time, for certain positions in aligned sample of transcription factor binding sites, it is possible to find a set of conformational and physicochemical properties that would remain constant for all the variants of sites despite difference in the context. This is first determined by the specificity of DNA–protein interactions for a particular DNA–protein complex (Oshchepkov *et al.*, 2004). These properties will have close values in all the variants of sites of the sample differing in their contexts, and thus, the analysis directed to search for variations in properties of the site will detect their low variance at particular positions within the site. Consequently, the complete set of data on the conservative conformational and physicochemical properties of sites reflect the specificity of DNA interaction with a particular protein and may be effectively used for recognition of potential binding sites.

Methods and Algorithms

The score value for SITECON algorithm (see Oshchepkov *et al*., 2004 for reference) corresponds to the probability of the properties of the DNA sequence analyzed to be close to the detected conservative properties of the sequences forming the learning set. Let us designate this value as the level of required conformational similarity or, in other words, this value is considered to be a "score" value and is compared with the particular "threshold" value to decide whether this sequence could be a "site" or "not site". For each threshold value we calculate correlation coefficient CC and averaged conditional probability ACP (Bajic *et al.*, 2000). This values in our opinion are the most adequate for estimation of prediction quality:

$$CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \quad (1)$$

$$ACP = \frac{1}{4}\left(\frac{TP}{TP + FN} + \frac{TP}{TP + FP} + \frac{TN}{TN + FP} + \frac{TN}{TN + FN}\right), \quad (2)$$

where TP (True Positives) – correct predictions made by the program in the training set,

FN (False Negatives) – number of sites that are not recognised as the sites in the training set,

FP (False Positives) – number of site predictions made in the random sample,

TN (True Negatives) – number of true non-site predictions in the random sample.

TP and FN parameters are calculted using the jack-knife method with exclusion of 1 of sequences in series from the training sample. TN and FP parameters were estimated by recognition of binding sites in a randomly generated sequence. The negative sequences were generated by random shuffling of nucleotides in the initial site sequences; thus, the nucleotide compositions of both the positive and negative samples were identical. The recognition was carried out in both directions.

Implementation and Results

Web resource SITECON contains the knowledge base of conservative properties for over 100 samples of sites presented in the the main menu, and allows for recognition of potential transcription factor binding sites basing on this data from both the knowledge base and results of analysis of a sample proposed by an user. In this resource, high quality representative samples constructed using the information compiled in the TRRD database (Kolchanov *et al*., 2002) are compiled. A site was added to the corresponding sample only if either its functionality was confirmed experimentally or binding of the transcription factor to the site was demonstrated by one of the following methods: EMSA with nuclear extract and specific antibodies, EMSA with purified or recombinant protein, DNase I footprinting with purified or recombinant protein, or trans-activation of a reporter gene by overexpression of a transcription factor together with mutation analysis of site. In addition, we used in this resource the samples constructed of artificially selected sequences binding to transcription factors with a high affinity from the database TRRD-ArtSite.

We have collected the data concerning recognition quality for all the samples available on the site. For each sample the recognition errors, CC and ACP values where calculated for different threshold values. Maximum CC and ACP values for each sample are presented in the Table. The samples are divided into two groups: natural binding sites and artificially selected sequences. For each group of binding sites, rather high values both for ACP and CC are demonstrated.

Table. Maximum CC and ACP values for TFBS available for recognition at SITECON web-site

Factor name	max. CC	max. ACP	Factor name	max. CC	max. ACP
AP1	0,515	0,770	AHR2 ARNT TRRD-ArtSite (AS00006)	0,832	0,916
ARNT	0,391	0,778	ALF2sel TRRD-ArtSite (AS00308)	0,659	0,845
CEBP A	0,456	0,746	AR TRRD-ArtSite (AS00323)	0,727	0,864
CEBP all	0,542	0,775	AREB6 TRRD-ArtSite (AS00030)	0,528	0,791
CLOCK	0,693	0,850	BCL6 TRRD-ArtSite (AS00242)	0,800	0,900
cMyc canonical	0,756	0,881	BRN3A TRRD-ArtSite (AS00209)	0,909	0,955
CRE	0,600	0,833	CF1 TRRD-ArtSite (AS00150)	0,718	0,868
CREB	0,361	0,761	cFOS TRRD-ArtSite (AS00008)	0,631	0,819
CREB zag,	0,379	0,757	CMYC_MAX2 TRRD-ArtSite (AS00129)	0,913	0,958
E2F/DP	0,738	0,870	CREB TRRD-ArtSite (AS00246)	0,612	0,812
EGR1	0,414	0,739	E2F1DP1 TRRD-ArtSite (AS00123)	0,562	0,790
ELKF	0,267	0,748	E2F1DP2 TRRD-ArtSite (AS00124)	0,732	0,870
ER2	0,516	0,783	E2F4DP1 TRRD-ArtSite (AS00125)	0,859	0,930
GATA all	0,509	0,782	E2F4DP2 TRRD-ArtSite (AS00126)	0,721	0,868
GATA1	0,557	0,783	E74A TRRD-ArtSite (AS00170)	0,341	0,778
GATA2	0,280	0,764	ELK1 TRRD-ArtSite (AS00270)	0,562	0,796
GATA3	0,251	0,756	ERRA TRRD-ArtSite (AS00268)	0,717	0,859
HNF1	0,402	0,736	FLI1 TRRD-ArtSite (AS00202)	0,810	0,906
HNF3,4	0,285	0,749	GATA1 TRRD-ArtSite (AS00058)	0,678	0,839
HNF4	0,473	0,762	GATA2 TRRD-ArtSite (AS00036)	0,641	0,825
IRF	0,370	0,751	HEN1 TRRD-ArtSite (AS00199)	0,918	0,959
ISRE	0,315	0,748	HFH8 TRRD-ArtSite (AS00105)	0,283	0,756
MyoD	0,385	0,737	HLF TRRD-ArtSite (AS00251)	0,866	0,933
Nf/E2	0,603	0,808	HNF1A TRRD-ArtSite (AS00106)	0,648	0,825
NFAT	0,522	0,778	HOX11 CTF1 TRRD-ArtSite (AS00344)	0,839	0,920
NFkB all	0,708	0,856	HOXA11 TRRD-ArtSite (AS00099)	0,751	0,883
NFkB hetero	0,574	0,791	IBR TRRD-ArtSite (AS00336)	0,902	0,952
NFkB homo	0,694	0,848	IKAROS1 TRRD-ArtSite (AS00152)	0,889	0,945
NRF2	0,711	0,866	IRF1 TRRD-ArtSite (AS00062)	0,709	0,858
OCT all	0,245	0,743	IRF2 TRRD-ArtSite (AS00063)	0,814	0,907
OCT1	0,218	0,745	LHX3 TRRD-ArtSite (AS00333)	0,470	0,768
P53	0,596	0,822	MAD MAX TRRD-ArtSite (AS00225)	0,712	0,857
Pu1	0,613	0,807	MAX2 TRRD-ArtSite (AS00131)	0,849	0,926
SF1	0,646	0,824	MEIS1A TRRD-ArtSite (AS00095)	0,654	0,828
SRE canonical	0,439	0,763	MIDA1 TRRD-ArtSite (AS00338)	0,693	0,850
SRF all	0,629	0,830	MOK2 TRRD-ArtSite (AS00195)	0,833	0,917
STAT	0,558	0,780	MYOG TRRD-ArtSite (AS00115)	0,809	0,905
STAT1	0,696	0,848	NFE2 TRRD-ArtSite (AS00034)	0,915	0,958
TTF1	0,257	0,740	NMYC TRRD-ArtSite (AS00253)	0,625	0,815
USF	0,622	0,816	OCT2 TRRD-ArtSite (AS00107)	0,937	0,969
YY1	0,290	0,750	PAX3 TRRD-ArtSite (AS00183)	0,881	0,941
			PAX6 TRRD-ArtSite (AS00009)	0,706	0,854
			PPARG TRRD-ArtSite (AS00023)	0,964	0,982
			PRD TRRD-ArtSite (AS00013)	0,935	0,969
			PRD TRRD-ArtSite (AS00014)	0,965	0,983
			RORA1 TRRD-ArtSite (AS00032)	0,758	0,894
			RORA2 TRRD-ArtSite (AS00033)	0,617	0,811
			RTR TRRD-ArtSite (AS00244)	0,937	0,969
			SAP1 TRRD-ArtSite (AS00285)	0,760	0,886
			SOX9 TRRD-ArtSite (AS00027)	0,722	0,869
			SREBP1 TRRD-ArtSite (AS00164)	0,787	0,897
			SRF TRRD-ArtSite (AS00007)	0,993	0,996
			STE11 TRRD-ArtSite (AS00065)	0,753	0,881
			TAX CREB TRRD-ArtSite (AS00247)	0,728	0,874
			TH1_E47 TRRD-ArtSite (AS00169)	0,550	0,782
			UNC86 TRRD-ArtSite (AS00211)	0,930	0,965
			USF TRRD-ArtSite (AS00239)	0,800	0,900
			USP ECR TRRD-ArtSite (AS00328)	0,702	0,854
			VJUN TRRD-ArtSite (AS00213)	0,948	0,974
			YY1 TRRD-ArtSite (AS00001)	0,925	0,962
			YY1 TRRD-ArtSite (AS00001)	0,925	0,962

Discussion

The Web resource SITECON, described in this work, provides detection of data on conservative properties of sites basing on statistical analysis of samples of transcription factor binding sites and contains data on 38 conformational and physicochemical DNA properties as well as allows for construction of recognition rules and search for potential transcription factor binding sites in genomic sequences. This resource also contains the knowledge base of conservative conformational and physicochemical properties of over 100 transcription factors. Along with the samples of transcription factor binding sites proposed by SITECON, a user may analyze the own data in a form of aligned sample of a transcription factor binding sites. High quality recognition demonstrated for samples proposed. Maximum ACP and CC values are higher in average for artificially selected sequences then for natural binding sites, this could be a consequence of similar mechanism of artificial selection for the sequences of the set.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grants Nos. 03-04-48469-a 02-07-90355, 03-07-90181-v, and 02-07-90359); Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); R A S Presidium Program "Molecular and Cellular Biology" (grant No. 10.4); and NATO (grant No. LST.CLG.979816).

References

Bajic V.B. Comparing the success of different prediction software in sequence analysis: a review // Brief. Bioinform. 2000. V. 1(3). P. 214–228.

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. V. 30. P. 312–317.

Meierhans D., Sieber M., Allemann R.K. High affinity binding of MEF-2C correlates with DNA bending // Nucleic Acids Res. 1997. V. 25. P. 4537–4544.

Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. SITECON—A tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition // Bioinformatics of genome regulation and structure / Ed. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London, 2004. P. 93–102.

Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Podkolodny N.L., Savinkova L.K., Kolchanov N.A., Overton G.C. Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins // Bioinformatics. 1999. V. 15. P. 687–703.

Starr D.B., Hoopes B.C., Hawley D.K. DNA bending is an important component of site-specific recognition by the TATA binding protein // J. Mol. Biol. 1995. V. 250. P. 434–446.

Suzuki M., Amano N., Kakinuma J., Tateno M. Use of 3D structure data for understanding sequence-dependent conformational aspects of DNA // J. Mol. Biol. 1997. V. 274. P. 421–35.

A DATABASE ON DNA SEQUENCE/ACTIVITY RELATIONSHIPS: APPLICATION TO PHYLOGENETIC FOOTPRINTING

Ponomarenko M.P.*, Ponomarenko J.V.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: pon@bionet.nsc.ru

Keywords: *transcription factor binding site, phylogenetic footprint, sequence-activity relationship, annotation, database*

Summary

Motivation: ACTIVITY, a database on DNA site sequences with known sequence-activity relationships under fixed experimental conditions has also been adapted to applications that perform phylogenetic footprinting of known sites.

Results: Three ACTIVITY_Reports, ACTIVITY_Tuning, and TFsite_Annotations resources have been developed. For a given site with known sequence-activity relationships, the ACTIVITY_Reports accumulates the quantitative data on the impact of the site's surroundings, which allows correct recognition of the site. The ACTIVITY_Tuning stores the Java script applets generated for the annotation of phylogenetic footprints of the sites with known sequence-activity relationships. TFsite_Annotations documents the annotation results.

Availability: http://util.bionet.nsc.ru/databases/activity.html

Introduction

In our paper (Ponomarenko *et al*., 2001) we introduced the ACTIVITY database accumulating DNA/RNA sequence-activity relationships experimental data. Since such data are condition-dependent, the question is how to use sequence-activity relationships obtained in one experiment to predict them in another experiment. Providing cross-validation tests between data obtained through different experimental conditions, we have shown that, for a given site, the impact of the site core into sequence-activity relationships is invariant, whereas relationships between the site and its surroundings could be condition-dependent and lead to varying activity values.

In recent years, quantitative measurement technological advancements in microarray analysis and SNP-disease association have focused the computation-based study of transcription factor binding DNA sites (TF sites) on their invariant positioning, mutual location, and surroundings. A number of recently developed approaches have taken into account the anchor locations of the key TF sites in the tissue specific promoters (Werner *et al*., 2003) and potential oligonucleotide-targets for many TF's synergistic specific to a given TF site (Kel *et al*., 2001). Also, the phylogenetic footprint approach has been adapted to the aim of the discovery of putative TF sites through interspecies genome comparison (Wasserman *et al*., 2000). Finally, statistical criteria for examining the invariant potential TF sites specifically clustered within the homologous, co-regulated and co-expressed gene promoters have been developed (Sosinsky *et al*., 2003).

Following the novel tendency in site-surrounding analysis, we have adapted ACTIVITY to the case of phylogenetic footprints of known TF sites. Accounting the complexity of molecular complexes of a cell, particularly eukaryotic transcription complex, the prediction of influence of local nucleotide surrounding of a DNA site to its binding with a target protein(s) is very challenged task. However, we can suggest that sites of the same type and at the same location within the same regions of homologous genes have the same impact of surrounding nucleotides to observed sequence-activity relationships due to common evolution origin. Thus, our recent developments have been focused on phylogenetic footprinting of protein-binding sites with known sequence-activity relationships.

Methods and Algorithms

Three resources have been developed in this work on the basis of our method and algorithm developed earlier (Ponomarenko *et al*., 2002). For a given site with known sequence-activity relationships, the database ACTIVITY_Reports accumulates the quantitative data on the impact of the site's surroundings, accounting for the possibilities of many transcription factors binding to the site. From a selected entry of this database, a Java script applet aimed at further precise analysis of phylogenetic footprints of the site is stored within the ACTIVITY_Tuning. The applet is used for the annotation of the site in homologous genes considering the site locations. The TFsite_Annotations documents the annotation results.

Figure represents an example entry of the TFsite_Annotations. The entry summarizes the results of annotation of the heat shock factor (HSF) binding element in the promoter regions of a number of genes. The annotation was obtained using the mutagenesis data in the human hsp70 promoter (Tsutsumi-Ishii *et al*., 1995), revealing that nucleotide substitutions -113A>G, -103A>G, -101T>A, and -98T>C decreased the transcription activity of the promoter to 40 %, 10 %, 90 %, and 60 %, of the wild type promoter activity, accordingly. The algorithm developed earlier (Ponomarenko *et al*., 2002) was used to analyze this data. It revealed the HSF binding to the –113/-98 bp region of human hsp70 promoter, which is consistent with the CAT-assay data (Tsutsumi-Ishii *et al*., 1995). The results are documented in the ACTIVITY_Reports, from which, upon receiving a report-entry, a Java script applet is stored in the ACTIVITY_Tuning. These databases are then used for the annotation of DNA sites in regulatory regions of homologous genes. A researcher analyzing arbitrary sequences by interest could also use the applets stored in the ACTIVITY_Tuning.

Implementation and Results

As an example of the implementation, forty-one genes homologous to the human hsp70 gene have been found in the EMBL Data Library. Only nine of them are documented as genes containing the HSF binding site. The sequence fragments of the promoters of these genes around position -113 relative to transcription start site are shown in the central part of the TFsite_Annotations entry (Fig.). The right part of the entry represents the significance levels for each specific rate of the HSF/DNA-binding considered. The results show that all nine known HSF binding sites have been annotated correctly because the events of site presence had higher significance than those of the site absence. The HSF regulates the expression of genes participating in heat-shock response via the heat shock elements (HSE). It is known that HSE is composed of 3-8 inverted repeats of the pentamer nGAAn. The sequence nGAAnnTTCnnGAAn has been determined as the "ideal" HSE consensus sequence (Lardans *et al*., 2001). However, some HSEs differ from "ideal", for example HSE of S.cerevisiae MDJ1 promoter has a sequence nTTCn-(11bp)-nGAAn-(5bp)-nGAAn (Tachibana *et al*., 2002). Among the HSEs predicted in this work and confirmed experimentally, there are HSEs which are not "ideal" HSEs (non-"ideal" bases are given in bold): (i) aaggtcataGAAagTTCtaG**t**Ac (Schistosoma mansoni, hsp70); (ii) ccGAAacT**g**CtgGAAgaTTCt (rat, hsp70); (iii) gcGAAac**c**TCtgGAAtaTTCcc (pig, hsp70.2); and (iv) a**ac**Cct**cg**AgcTTCat**c**TCaa (carrot hsp70). Case iv is particularly exciting, where the HSE consensus has not been found, but HSE has been successfully annotated in consistence with that has been experimentally found.

Currently, the annotation of 155 TF sites in the promoters of 101 eukaryotic genes can be found in the ACTIVITY database. Of these sites, 142 sites are not yet documented.

Discussion

As the discussible implementation and results, TF site predictions for many non-experimental putative genes are also provided. They helpful in planning sequence-activity experiments.

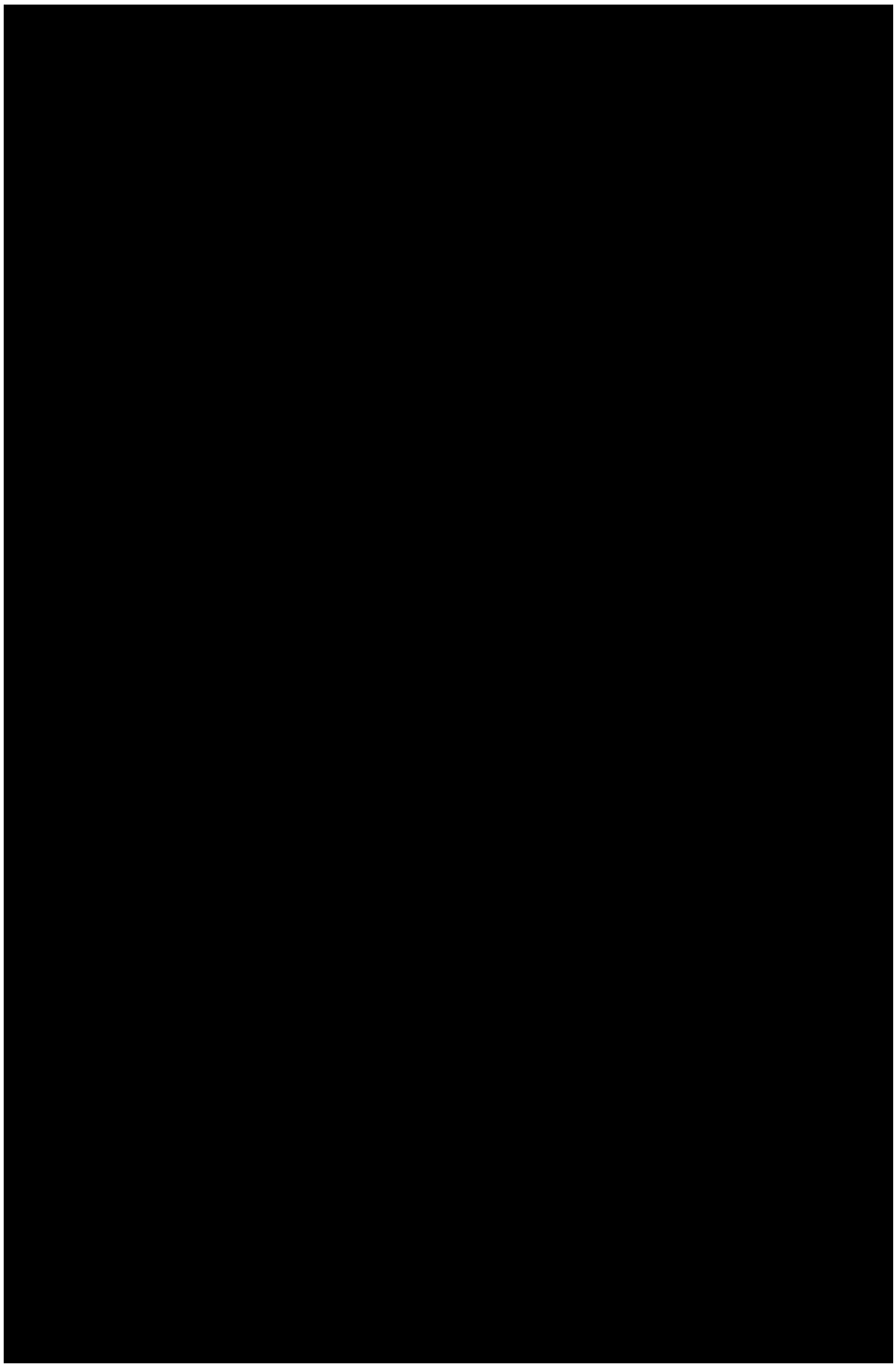


Fig. An example entry of the TFsite_Annotations, http://util.bionet.nsc.ru/databases/activity_annotation_m000_list.html.

References

Kel A., Kel-Margoulis O., Farnham P., Bartley S., Wingender E., Zhang M. Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors // J. Mol. Biol. 2001. V. 309. P. 99–120.

Lardans V., Ram D., Lantner F., Ziv E., Schechter I. Differences in DNA-sequence recognition between the DNA-binding domain fragment and the full-length molecule of the heat-shock transcription factor of schistosome // Biochim. Biophys. Acta. 2001. V. 1519. P. 230–234.

Ponomarenko J., Furman D., Frolov A., Podkolodny N., Orlova G., Ponomarenko M., Kolchanov N., Sarai A. ACTIVITY: a database on DNA/RNA sites activity adapted to apply sequence-activity relationships from one system to another // Nucleic Acids Res. 2001. V. 29. P. 284–287.

Ponomarenko J., Merkulova T., Orlova G., Gorshkova E., Fokin O., Ponomarenko M., Frolov A., Sarai A. Mining DNA sequences to predict sites which mutations cause genetic diseases // Knowledge-Based Systems. 2002. V. 15. P. 225–233.

Sosinsky A., Bonin C., Mann R., Honig B. Target explorer: an automated tool for the identification of new target genes for a specified set of transcription factors // Nucleic Acids Res. 2003. V. 31. P. 3589–3592.

Tachibana T., Astumi S., Shioda R., Ueno M., Uritani M., Ushimaru T. A novel non-conventional heat shock element regulates expression of MDJ1 encoding a DnaJ homolog in *Saccharomyces cerevisiae* // J. Biol. Chem. 2002. V. 277. P. 22140–22146.

Tsutsumi-Ishii Y., Tadokoro K., Hanaoka F., Tsuchida N. Response of heat shock element within the human HSP70 promoter to mutated p53 genes // Cell Growth Differ. 1995. V. 6. P. 1–8.

Wasserman W., Palumbo M., Thompson W., Fickett J., Lawrence C. Human-mouse genome comparisons to locate regulatory sites // Nat. Genet. 2000. V. 26. P. 225–228.

Werner T., Fessele S., Maier H., Nelson P. Computer modeling of promoter organization as a tool to study transcriptional coregulation // FASEB J. 2003. V. 17. P. 1228–1237.

ANALYSIS OF GENE REGULATORY SEQUENCES BY KNOWLEDGE DISCOVERY METHODS

***Pozdnyakov M.A.*[1], *Orlov Yu.L.*[1], *Vishnevsky O.V.*[1], *Proscura A.L.*[1], *Vityaev E.E.*[2], *Arrigo P.*[3]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia, e-mails: {mike,orlov,oleg,anya}@bionet.nsc.ru; [2] Sobolev Institute of Mathematics, Novosibirsk, 630090, Russia, e-mail: vityaev@math.nsc.ru; [3] ISMAC, via De Marini 6 16149 Genova, Italy, e-mail: arrigo@ge.ismac.cnr.it

* Corresponding author: e-mail: mike@bionet.nsc.ru

Keywords: *eukaryotic promoter, recognition, transcription factors binding sites, Machine Learning, Knowledge Discovery, Data Mining*

Summary

Motivation: Analysis of gene regulatory sequences is of great interest for understanding molecular mechanisms of gene expression. We present implementation of Knowledge Discovery techniques for regularities search in tables of context features of DNA sequences involved in gene transcription regulation.

Results: Discovery of regularities is based on construction of complex signals of these features including experimental information on gene expression and prediction of potential transcription factor binding sites. The search patterns for regularities have been constructed in the first-order logic with probabilistic estimates. The sets of promoter nucleotide sequences were selected from TRRD and EMBL databases based on tissue-specificity, type of gene regulation or joint expression in a functional system. Set of regularities relating the nucleotide sequences and gene functional class were found. Prediction by regularities of tissue-specific promoters in long genomic DNA showed high accuracy.

Availability: the software is available by request to the author vityaev@math.nsc.ru

Introduction

The method presented discovers the functional type of regulatory region based on a predicted set of protein binding sites and the context signals in the nucleotide sequence (Vityaev *et al*., 2001; Kolchanov *et al*., 2003). Computational approaches for understanding the transcriptional regulatory network include promoter prediction and combinatorial regulatory elements prediction (Qiu, 2003). The main goal of this research is to identify gene function by using a set of integrated methods for the recognition of regulatory elements and transcription factors binding sites.

Initially set of context sequence features and potential (predicted) TFBS need to be constructed for contrast training sample of promoter sequences. Our approach include: 1) Computer-assisted discovery of potential binding sites within the sequence under analysis (consensus, weight matrices, prediction by homology with databases); 2) Search of specific oligonucleotides, low complexity regions and other context features; 3) Constructing of table "object-character" for the sample of sequences. Then, the system analyzes combinations of joint presence of features in sequence.

Distinctive feature of our approach is the usage of specific feature patterns describing a subgroup of the training set. (Vityaev *et al*., 2001) The search patterns for regularities are constructed in the first-order logic augmented by probabilistic estimates. The program is written in the C++ and it is supplied by a user-friendly interface.

We analyzed the sets of promoter nucleotide sequences selected from TRRD and EMBL databases by tissue-specificity and type of gene regulation. Set of regularities relating the nucleotide sequences and gene functional class were found. Prediction by regularities of tissue-specific promoters in long genomic DNA showed high accuracy.

Systems and Methods

Data. Promoter sequences were extracted from TRRD database (Kolchanov *et al*., 2002) and divided into several groups according to the tissue-specificity (promoters of endocrine system genes, cholesterol homeostasis, heat shock response system, interferon-regulated, glucocorticoid-regulated, cell-cycle system).

Methods. Computer system "Gene Discovery" consists of three main modules: 1) the module for on-line representation of the context signals from DNA sequence in a standard table form; 2) the module "Discovery" for regularities searching; (3) the module for recognition of the sequence class by using the regularities found.

System "Discovery", retrieve statistically significant first-order logic rules for functional annotation of regulatory regions. This system base on first-order representations of data and hypotheses (Vityaev *et al*., 2001). As with any technique based on logic rules, this technique allows one to obtain human-readable forecasting rules and provides promoter recognition.

The signal in nucleotide sequence could be: 1) potential functional site, predicted by homology (or weight matrix) with annotated sequences in the specialized molecular-biology database; 2) context signal (specific oligonucleotide); 3) site with conserved conformational or physical-chemical features (such as double-helix angles, DNA melting temperature); 4) secondary structure element (low complexity region, Z-DNA, RNA hairpin).

An example of formal TFBS presentation as a signal is given in the scheme below:

```
-------------------------------------------------
<Matrics_TFBS>
      Signal_Number 272
      <Signal 1>
   name 103_AP2_AS00103
   TAGAAAGCCCCGGT
            method_name weight_matrics
      </Signal 1>
```

All these signals may be recognized using knowledge about DNA properties and the consensus scheme based on experimental data stored in specialized databases.

Resulting regularities are stored in the software as complex signals in IF-THEN form

$$(A_1 \& \ldots \& A_k) \Rightarrow A_0,$$

where the IF-part, $A_1 \& A_2 \& \ldots \& A_k$, consists of true/false logical statements $A_1,\ldots,A_k$, concerning presence of context features (potential TFBS) in a sequence and the THEN-part consists of a single logical statement A_0, concerning promoter class and gene function.

Example of a pattern of potential binding sites location in endocrine system gene promoters is given in Figure 1. The promoter sequences are aligned relative to the transcription start (position +1 bp), indicated by arrows. Gene names are given at the left.

Potential transcription factor binding sites composing the complex signal are shown as gray rectangles.

The great number of regularities for joint appearance of the context signals in the promoter regions was found by "Gene Discovery" system. Recognition rules are based on complex signals found. Profile of promoter prediction function for long genomic DNA is shown in Figure 2. Promoter sequences from control sample are distributed periodically in the sequence. Higher peaks of profile correspond to correct promoter prediction (Fig. 2).

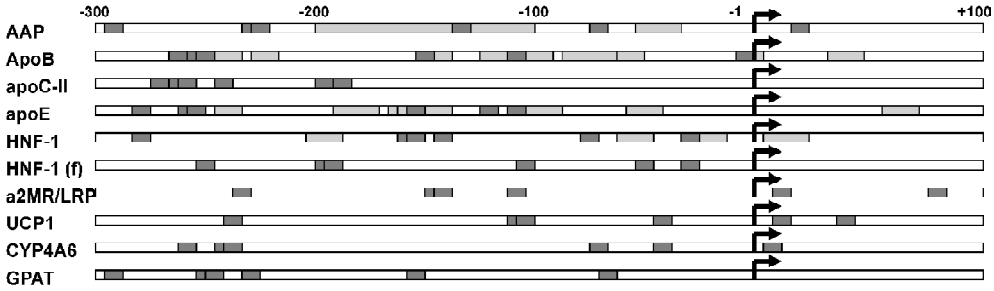


Fig. 1. An example of pattern of potential binding sites location in promoter sequences (grey rectangles). Gene identifiers: human AAP (A00596), human ApoB (A00149), human apoC-II (A00350), human apoE (A00151), rat HNF-1 (A00161), clawed frog HNF-1 (A00382), human alpha2MR/LRP (A00417), rat UCP1 (A00915), rabbit CYP4A6 (A00640), mouse GPAT (A00414).

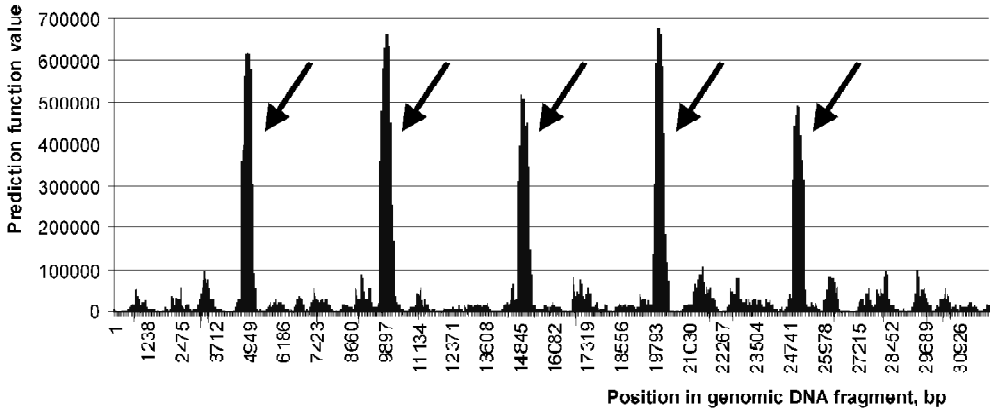


Fig. 2. Profile of promoter prediction function based on complex signals. Arrows indicate correct promoter location in control sequence (simulated genomic DNA by the length 100Kb). Control sequence includes cholesterol homeostasis gene promoters.

Results and Discussion

Thus, the system "Gene Discovery" helps us to find complex signals in promoter regions. In a similar way any samples of phased nucleotide sequences could be analyzed. The functional meaning of the signal could be treated in terms of the transcription factors binding sites or the conformational properties of DNA.

Distinctive features of the approach suggested are: 1) combination of rules not restricted by fixed order (like only pairs or triples of signals); 2) using original method for context feature selection like oligonucleotides, weight matrices from TRRD, TRANSFAC and SELEX databases, nucleosome formation estimation, nucleotide secondary structure signals, low complexity regions; 3) using verified experimental data on transcription factor binding site location from TRRD database.

The regularities found could be analyzed by a molecular biology expert as unique complex signals, which are significant for proper promoter functioning. The research suggested that functional promoter modules could be detected by formal models independent of homology level between sequences.

This Data Mining approach is applicable for analysis of context gene structure at all levels of gene hierarchy: promoter, regulatory regions, transcription factor binding sites (Liu, Wong, 2003). The algorithm has a flexibility to search for structural patterns that are typical for a whole set of sequences as well as for a subset of sequences.

Acknowledgements

The authors are grateful to N.L. Podkolodny and N.A. Kolchanov for scientific discussion. This work was supported in part by the Russian Foundation for Basic Research (02-07-90355, 03-04-48506), Ministry of Education (E02-6.0-250), NATO (LST.CLG 979815), Integration project of SB RAS (119), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology", Russian Ministry of Industry, Sciences and Technologies Grant No. 43.073.1.1.1501.

References

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription Regulatory Regions Database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. V. 30(1). P. 312–7.

Kolchanov N.A., Pozdnyakov M.A., Orlov Yu.L., Vishnevsky O.V., Podkolodny N.L., Vityaev E.E., Kovalerchuk B. Computer System "Gene Discovery" for Promoter Structure Analysis // Artificial Intelligence and Heuristic Methods in Bioinformatics / Eds. P. Frasconi, R. Shamir. IOS Press, 2003. P. 173–192.

Liu H., Wong L. Data mining tools for biological sequences // J. of Bioinformatics and Computational Biol. 2003. V. 1(1). P. 139–167.

Qiu P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network // Biochem Biophys Res Commun. 2003. V. 309(3). P. 495–501.

Vityaev E.E., Orlov Yu.L., Vishnevsky O.V., Belenok A.S., Kolchanov N.A. Computer system Gene Discovery for regularities search in eukaryotic regulatory regions // Mol. Biol. (Mosk.). 2001. V. 35(6). P. 952–960. (in Russian).

SREBP BINDING SITES: CONTEXT FEATURES AND ANALYSIS OF GENOME DISTRIBUTION BY THE SITECON METHOD

Proskura A.L., Oshchepkov D.Yu., Pozdnyakov M.A., Ignatieva E.V.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: å-mail: eignat@bionet.nsc.ru

Keywords: *database, LM-TRRD, transcription regulation, lipid metabolism, SREBP*

Summary

Motivation: Disruption of lipid metabolism is known to cause a set of severe human diseases. Transcription factor SREBP (Sterol Regulatory Element Binding Protein) is a key regulator of cholesterol homeostasis gene expression, hence, analysis of SREBP binding site data, as well as development of a reliable method for SREBP binding sites recognition are extremely important tasks.

Results: We have performed a context analysis of two subsamples of transcription factor binding sites (TFBS) SREBP: SRE and E box. For these two types of TFBS, we have found essential differences in their context organization. We have developed a method SITECON aimed at recognition of the sites referring to the SRE type. By the method SITECON, we have studied distribution of potential TFBS of the SRE type in various genome sequences. As found, promoter regions of cholesterol homeostasis are characterized by higher concentration of the SRE sites than promoter regions of the other genes. Also, exons of the lipid metabolism genes are better saturated with potential SRE sites than exons of other groups of genes.

Introduction

Transcription factors of the SREBP family play an important role in regulation of expression of genes controlling cholesterol level and synthesis of triglycerides in a cell. The active SREBP form is obtained from inactive precursor, this process being suppressed by increasing inner cellular cholesterol level (Brown, Goldstein, 1997). Expression of SREBP factors is activated by insulin and carbohydrates and suppressed by fatty acids (Osborne, 2000). As known, the factors of this family, SREBP1a, SREBP1c, and SREBP2, belong to the family of bHLHLZ (basic helix-loop-helix leucine zipper) proteins and bind to the sites like E-box and non-E-box (SRE). The sites of E-box type are formed by a palindrome (ATCACGTGA) consisting from two half-sites, 5‘-TCAC-3 and 5‘-GTGA-3. The sites of the SRE type (Sterol Regulatory Element) (ATCACCCCAC) are represented by imperfect repeat 5‘-ATCAC-3‘ and 5‘-CCCAC-3‘ (Kim *et al.*, 1995). In the present study, we perform a context analysis of two sub-types of SREBP sites. We have also studied the composition of potential sites of this type, located both in regulatory regions of cholesterol homeostasis genes and of various functional groups.

Methods

Samples of TFBS of regulatory gene regions, exons, and introns. The samples of SREBP TFBS (53 sequences) and promoter regions of genes referring to various functional groups were automatically generated from the TRRD (http://www.bionet.nsc.ru/trrd/) and EMBL nucleotide sequence database. We have divided the SREBP site sample into two sub-samples: SRE (38 sites) and E-box type (15 sites). The sample of 8 genes of cholesterol homeostasis, where SREBP sites are not detected yet experimentally, includes the genes: human MSR, SRB1, SCAP, ABCG1, INSIG1, S1P, mouse CAV1, and chicken FAS. The set of promoter regions of human genes was extracted from the database EPD (http://www.epd.isb-sib.ch/seq_download.html) (1871 sequences with the length (-300/+100)). The samples of 209 introns, 220 coding and 72 non-coding exons and

8 3‘-flanking regions for more than 50 genes of the lipid metabolism system were extracted from the EMBL nucleotide sequence database. The total length of sequences in these samples equals to 176122, 28348, 22813, and 13720 nucleotides, respectively. The samples of exons (3 samples with the total length 512967 bp) and introns (2 samples with the total length 9149111 bp) were extracted from the database accumulating exons-introns (http://mcb.harvard.edu/gilbert/EID/).

Matrices, estimation of $\chi 2$. By analyzing the SREBP sample (53 sequences), we have constructed the frequency matrix of the SREBP site, including SRE and E-box sub-types (Table 2). We have evaluated the probability P of observed deviations of the base frequencies at each position from the expected ones due to random reasoning (Fig. 1).

Recognition of the SRE. SITECON is a method for recognizing sites based on analysis of their conservative physicochemical and conformational properties (Oshchepkov *et al.*, 2004). As a recognition threshold, the SITECON method employs the level of necessary conformational similarity, which was set on 4 levels for test SRE recognition quality. The recognition quality for type I errors was checked by the jack-knife method with removing 1 sequence in series from the training sample. The control for type II errors was performed by recognition of binding sites in the sequence, which was generated by random shuffling of nucleotides in the initial site sequences; thus, nucleotide composition of positive and negative samples was identical, and recognition was carried out in both directions. Evaluation of type I and II errors for the levels of conformational similarity 0.73, 0.74, 0.75, and 0.76 is shown in Table 1. Recognition of SRE was performed at the level of conformational similarity 0.74.

Table 1. Recognition errors for various levels of conformational similarity

	0.73	0.74	0.75	0.76
Type I error	0.5526	0.5526	0.6053	0.6316
Type II error	6.30E-03 (1/1586)	4.40E-04 (1/2271)	3.20E-04 (1/3123)	2.70E-04 (1/3701)

Results

Context analysis of SREBP binding sites. We have aligned nucleotide sequences in sub-samples of the SREBP sites (SRE-type and E-box type). To reveal the conserved region limits in both types of sites, we have constructed frequency matrices. Significance of deviation of nucleotide frequency was estimated and compared to monotonous distribution (0.25) in accordance with $\chi 2$ criterion (Fig. 1). In the sample of SRE, the conserved region was of nine nucleotides in length. The consensus TCASNCCAB derived from the SRE matrix (Table 2) corresponds to consensus found in literature in-between the second and the tenth nucleotides. In the sample of E-box sites, the conserved region was of six nucleotides in length (Fig. 1). The consensus derived from E-box matrix (CASSYK) corresponds to the central part of E-box ten-lettered consensus (Table 2). So, for the sites of E-box type, both the size of conserved region and the level of its conservatism estimated by $\chi 2$ criterion became essentially less than for the SRE. Probably, this fact is caused by a small sample size (15 sequences), and by increasing the sample size, the conserved nucleotides will be found in more extended region (Table 2).

Analysis of content of potential SRE in promoter gene regions (-300/-1) referring to various functional groups. The largest density of potential sites (2.83/1000 bp) was detected within –300/-1 bp regions of cholesterol metabolism genes (Fig. 2, 1st column). With least density, the sites were typically occurring in gene promoter regions compiled on the basis of total TRRD except genes of cholesterol homeostasis and other genes of lipid metabolism (Fig. 2, 6th column).

For the genes referring to other functional groups (columns 3, 4, and 6) and human genes extracted from the EPD (7th column), intermediate values of SRE density were detected. Also, we have evaluated the content of potential SRE in 8 genes controlling cholesterol homeostasis, in which SREBP sites were not found experimentally. The density of the potential SRE in this group (column 2) approximates to density of sites in a group of genes regulated by SREBP (column 1, Fig. 2).

1. 20 genes of cholesterol homeostasis
2. 8 genes of cholesterol homeostasis, where SREBP sites are not detected experimentally yet
3. 40 genes of erythroid system
4. 38 genes of lipid metabolism system, except the groups 1 and 2)
5. 75 genes of cell cycle
6. 352 genes from TRRD, except groups 1, 2, 3
7. 1871 human genes from EPD.

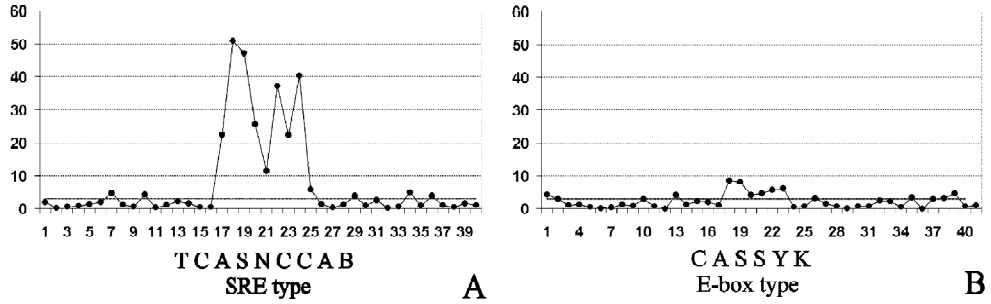


Fig. 1. The profile of –ln(P) values of significance level of $\chi 2$ statistics for frequency matrix positions of SRE (A) and E-box (B) types. By X axis, nucleotide position in a frequency matrix is given. Horizontal line marks -ln(0.05).

Table 2. Frequency matrix of the SREBP site

SRE										
Nucleotide	Positions									
	16	17	18	19	20	21	22	23	24	25
A	11	4	0	36	2	6	0	3	34	1
C	10	4	37	1	29	23	33	28	3	17
G	11	2	1	1	6	4	1	4	0	11
T	6	28	0	0	1	5	4	3	1	9
Consensus [1]	N	T***	C****	A****	S***	N**	C****	C***	A****	B*
Consensus [2]	A	T	C	A	C	C	C	C	A	C
E-box										
Nucleotide	Positions									
	16	17	18	19	20	21	22	23	24	25
A	6	2	0	11	1	1	0	1	5	3
C	6	5	11	1	9	4	5	1	5	2
G	2	2	3	2	3	9	1	10	3	4
T	1	6	1	1	2	1	9	3	2	6
Consensus [1]	M	Y	C****	A****	S*	S*	Y**	K**	V	D
Consensus [2]	A	T	C	A	C	G	T	G	A	T/C

* P<0.05, ** P<0.01 ***, P<0.001 ****, P<0.0001.

[1] Consensus is given in 15-lettered code, where **Y**-C,T; **M**-A,C; **K**-T,G; **S**-C,G; **B**-C,T,G; **V**-A,C,G.

[2] Consensus (Kim *et al*., 1995).

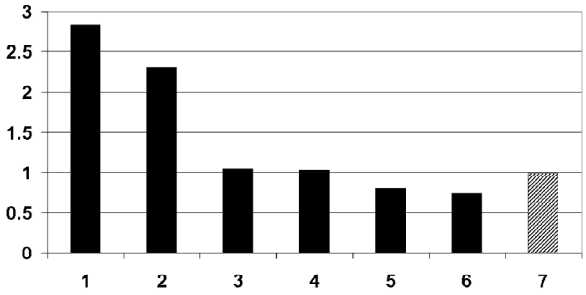


Fig. 2. Density of potential SRE recognized by SITECON method in promoter regions (-300/-1) of genes referring to various functional groups. Y axis, density of predicted sites per 1000 bp. Black columns, samples from TRRD, hatched column, a sample of genes extracted from EPD.

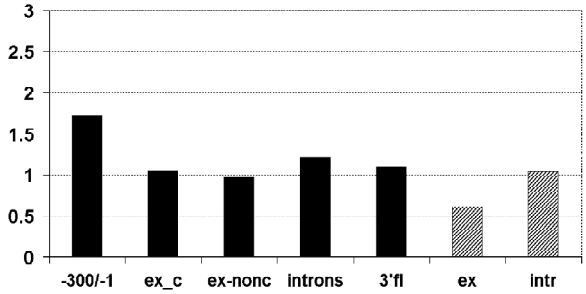


Fig. 3. Density of potential sites of SRE type detected by the SITECON method at the gene regions with different localization of sites from transcription start. Y axis, the number of predicted sites per 1000 bp. Black columns, regions of genes of lipid metabolism: 300/-1 promoter region; ex_c – coding exons; ex_nonc – noncoding exons; 3'fl – 3'-flanking regions. Hatched columns, exons and introns of the genes stored in Exon-Intron Database.

Analysis of potential SRE content within the regions of lipid metabolism genes, with different localization relatively transcription start. By using the SITECON method, we have analyzed the content of potential SRE in promoter regions (-300/-1), exons, introns, and 3'-flanking regions of lipid metabolism genes, including cholesterol homeostasis genes and others. The largest density of potential sites (1.72/1000 bp) was found in promoter (-300/-1) region (column 1). The lower content of potential sites (ranging from 0.9 to 1.21/1000 bp) was found in exons, introns, and 3'-flanking regions of lipid metabolism genes. For comparison, we have analyzed the samples of exons and introns from the Exon-Intron Database. The density of predicted sites in introns extracted from this database could be compared to the density in introns within the genes of lipid metabolism. In the sample of exons extracted from the Exon-Intron Database, the density of potential SRE sites was essentially lower than in all studied regions of lipid metabolism genes (Fig. 3).

Discussion

By analyzing the context of two SREBP site sub-types, we have revealed the difference in their structural organization and proved conception about structural organization of sites (Kim *et al.*, 1995). The model of the site of E-box type is the sequence (ATCACGTGA) including the palindrome TCACGTGA. The model of the site of SRE type is the sequence ATCACCCCAC including two copies of trinucleotide CAC. By taking into account the differences in the context organization of SREBP sites, it seemed reasonable to develop a recognition method for each of two sub-types individually. In this work, we present the results of testing of the SITECON method aimed at recognition of the SRE type sites. We have analyzed the density of potential SRE in

various genome sequences. The results obtained in general are in a good agreement with the expected distribution of SREBP sites in the groups studied, thus, giving the evidence about acceptable quality of the method developed. By applying SITECON method, we have demonstrated that the density of SRE in the group of cholesterol homeostasis genes, in which the SREBP sites were not experimentally detected yet, was close to that of genes regulated by SREBP (Fig. 3). So, the genes from this group are perspective for experimental testing.

Acknowledgements

The authors are grateful to Prof. Nikolay A. Kolchanov for the fruitful discussions, O.A. Podkolodnaya and I.I. Turnaev for kindly granted samples of regulatory gene regions. The work was supported in part by the Russian Foundation for Basic Research (Nos. 03-04-48506, 03-07-90181, 03-04-48469), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Project No. 119), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”.

References

Brown M.S., Goldstein J.L. The SREBP Pathway: regulation of Cholesterol metabolism by proteolysis of a membrane-bound transcription factor // Cell. 1997. V. 89. P. 331–340.

Kim J.B., Spotts G.D., Halvorsen Y.D., Shih H.M., Ellenberger T., Towle H.C., Spiegelman B.M. Dual DNA binding specificity of ADD1/SREBP1 controlled by a single amino acid in the basic helix-loop-helix domain // Mol. Cell. Biol. 1995. V. 15. P. 2582–2588.

Osborne T.F. Sterol Regulatory Element-binding Proteins (SREBPs): key regulators of nutritional homeostasis and Insulin action // J. Biol. Chem. 2000. V. 275. P. 32379–32382.

Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. SITECON – A tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition // Bioinformatics of genome regulation and structure / Ed. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/ London, 2004. P. 93–102.

APPLICATION OF TIME-FREQUENCY ANALYSIS IN EXON CLASSIFICATION

Renjun Yu.*, Eng Chong Tan

School of Computer Engineering, Nanyang Technological University, 639798, Singapore
* Corresponding author: e-mail: yurenjun@pmail.ntu.edu.sg

Keywords: *time-frequency distribution, exon, intron, digital signal processing*

Summary

Motivation: Digital signal processing (DSP) techniques are applied to the analysis of biological data.

Results: DSP techniques such as discrete Fourier transform and time-frequency distributions are implemented to analyze the exon and intron regions of DNA sequences by their time and or frequency properties. A proper map is discussed before DSP tools can be effectively applied. The HMR195 dataset (http://www.cs.ubc.ca/~rogic/evaluation/dataset.html) is used to evaluate the performance of the algorithms. For each sequence in HMR195, some other sequences from the same organism are used as training data.

Availability: MATLAB codes.

Introduction

A gene is a segment of DNA involved in producing a polypeptide chain. The coding regions are named as exons and the intervening sequences between exon regions are named as introns. There is structure and component difference between exon and intron regions. The focus of gene prediction research is to find the exon-intron structure of genes (Anthony *et al*., 1999). The DNA sequences can be viewed as character strings, and be digitalized into numerical sets (Coward, 1997). So DSP techniques can be applied for searching useful information buried in those sequences (Anastassiou, 2002).

Methods

(A) Discrete Fourier Transform (DFT)

Assume that the numbers a, t, c, and g are assigned to the characters A, T, C, and G, respectively. Then a DNA sequence of length N can be represented as follows:

$$x[n] = au_A[n] + tu_T[n] + cu_C[n] + gu_G[n],\ n = 0,\ 1,\ \ldots,\ N-1, \tag{1}$$

in which $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ are called the binary indicator sequences which take the value of either one or zero at location *n*, depending on whether or not the corresponding character exists at location *n*. Because only one nucleotide is possible at any index, the functions $u_A[n]$, $u_T[n]$, $u_C[n]$ and $u_G[n]$ add up to 1 for all *N*. Therefore any three of these four binary indicator functions are sufficient (Anastassiou, 2002). The DFT of a sequence $x[n]$ of length *N* is *X*[*n*], which is another sequence with the same length *N* given by (John, Dimitris, 1996):

$$X[n] = \sum_{n=0}^{N-1} x[n] e^{-j\frac{2\pi}{N}kn},\ k = 0,\ 1,\ \ldots,\ N-1. \tag{2}$$

It implies that the corresponding four DFT sequences are also a redundant set. The Fourier transform of (1) can be expressed as

$$W[k] = aA[k] + tT[k] + cC[k] + gG[k],\ k = 0,\ 1,\ 2,\ \ldots,\ N-1 \quad (3)$$

We can think of W as being a random variable and for properly chosen values of a, t, c and g, it is proved to be a superior predictor of whether or not a given DNA segment belongs to a coding region. We want to maximize the discriminatory capability between protein coding regions (with respective random variables A, T, C and G) and random DNA regions. Using a random number generator we can synthesize a DNA sequence of the same length of our statistical sample and therefore obtain the corresponding random variables A_R, T_R, C_R, G_R. Because the sequences A, T, C and G represent a redundant set, we can set one of the four coefficients to be zero, i.e.,

$$W = a \cdot A + t \cdot T + g \cdot G, \quad c = 0 \quad (4)$$

Therefore the optimization problem can be expressed as: "find the complex numbers a, t and g that maximize the quantity" (Anastassiou, 2002):

$$f(a,t,g) = \frac{E\left\{\left|aA + tT + gG\right|\right\} - E\left\{\left|aA_R + tT_R + gG_R\right|\right\}}{std\left\{\left|aA + tT + gG\right|\right\} + std\left\{\left|aA_R + tT_R + gG_R\right|\right\}}, \quad (5)$$

where std is the standard deviation under the constraints,

$$E\left\{\arg\left(aT + tT + gG\right)\right\} = 0,\ |a| + |t| + |g| = 1. \quad (6)$$

Using the optimization toolbox in Matlab, this optimization problem can be solved easily.

(B) Time-Frequency Distributions (TFDs)

The Wigner-Ville Distribution (WVD) at a particular time is obtained by (Cohen, 1995):

$$W(t,\omega) = \frac{1}{2\pi}\int s^*\left(t - \frac{1}{2}\tau\right)s\left(t + \frac{1}{2}\tau\right)e^{-j\tau\omega}d\tau. \quad (7)$$

However, the WVD suffers from the fact that cross-terms in multi-component signals bring confusing artifacts (Cohen, 1995). Two methods to improve results of WVD are: (i) the smoothing function that suppresses artifacts, and (ii) the reassigned method that provides a good concentration of signal (Auger, Flandrin, 1995). Combining the two methods, a more powerful TFD called the Reassigned Smoothed Pseudo WVD (RSPWVD) is obtained:

$$RSPWV_x(t',v';g,h) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} SPWV_x(t,v;g,h)\cdot\delta(t'-t''(x;t,v))\delta(v'-v''(x;t,v))dtdv. \quad (8)$$

Using a Gaussian function as the smoothing function, the smoothed pseudo WVD (SPWVD) is given by

$$SPWV_x(t,v;g,h) = \int_{-\infty}^{+\infty} h(\tau)\int_{-\infty}^{+\infty} g(s-t)\cdot x\left(s + \frac{\tau}{2}\right)x^*\left(s - \frac{\tau}{2}\right)dse^{-j2\pi v\tau}d\tau, \quad (9)$$

where x is the signal, h is the window function, g is the Gaussian Function, t is the time instant, v is the frequency instant and τ denotes the lag variable; t'' is the assigned time instant at which the center of gravity of energy contributions is located, and v'' is the assigned frequency instant at which the center of gravity of energy contributions is located, i.e.,

$$t'' = t - \frac{SPWV_x(t,v;\Gamma_g,h)}{2\pi\left[SPWV_x(t,v;g,h)\right]}, \quad v'' = v + \frac{SPWV_x(t,v;g,D_h)}{2\pi\left[SPWV_x(t,v;g,h)\right]}, \tag{10}$$

where $\Gamma_g(t) = t^* g(t)$ and $D_h(t) = \frac{d}{dt}[h(t)]$. Here we can use the frequency energy E to provide a similar result of DFT, i.e. $E = \int_{-\infty}^{+\infty} |TFD(t,\upsilon)| d\upsilon$.

(C) Detection Algorithm

The following algorithm is used to perform the exon classification work.

1. The HMR195 data in the form of character sequence is used as input for detection.
2. A set of DNA sequences known to from the same organism of each sequence in HMR195 is used as training data.
3. A set of random sequences of the same length of training data is generated for the computation of optimization parameters.
4. Four indicator sequences are used to represent DNA sequence.
5. TFDs and frequency energy are computed for each indicator sequence.
6. Optimization functions are used to get the optimization parameters.
7. Optimization parameters and the frequency energy of TFDs of the DNA data are used for computing the adapted frequency energy.
8. Peak detection is performed based on the frequency energy of the DNA data.

Results and Discussion

Using DFT, WVD and RSPWVD, F56F11.4a sequence (Genbank number: AF099922) is computed. Fig. shows the results of its subsequence with a length of 40,000 bases. From former research, this subsequence contains a gene with five exons. So it shows that all the three methods can identify these exons by peaks in the plot. Among these results, the results of WVD can display the exons more concentrated than DFT. The peak length in the result of WVD is shorter than that in DFT. The improvement in removing noise can easily be noted in the result of RSPWVD. We have tested 54 sequences of the HMR195 dataset. For each sequence we used similar sequences from the same organism as training data. The exon level sensitivity (*ESn*) and specificity (*ESp*) are defined as

$$ESn = \frac{TE}{AE}, \quad ESp = \frac{TE}{PE}, \tag{11}$$

where *TE* (true exons) is the number of exactly predicted exons and *AE* and *PE* are the numbers of annotated and predicted exons, respectively. Table shows the results. Using the simple techniques we are able to predict the exons locations in a stretch of DNA sequence. Among all the three methods, the RSPWVD provides the most satisfying results but the computation cost is the highest. The WVD also provides reasonably good accuracy but its noise can give rise to uncertain detections. The result of DFT is not as good as the RSPWVD but the results are acceptable too.

Table. Comparison of results

Methods	Detection results	Exon accuracy	
		ESn	ESp
DFT	54 (4)	0.67	0.73
WVD	54 (6)	0.63	0.68
RSPWVD	54 (2)	0.72	0.75

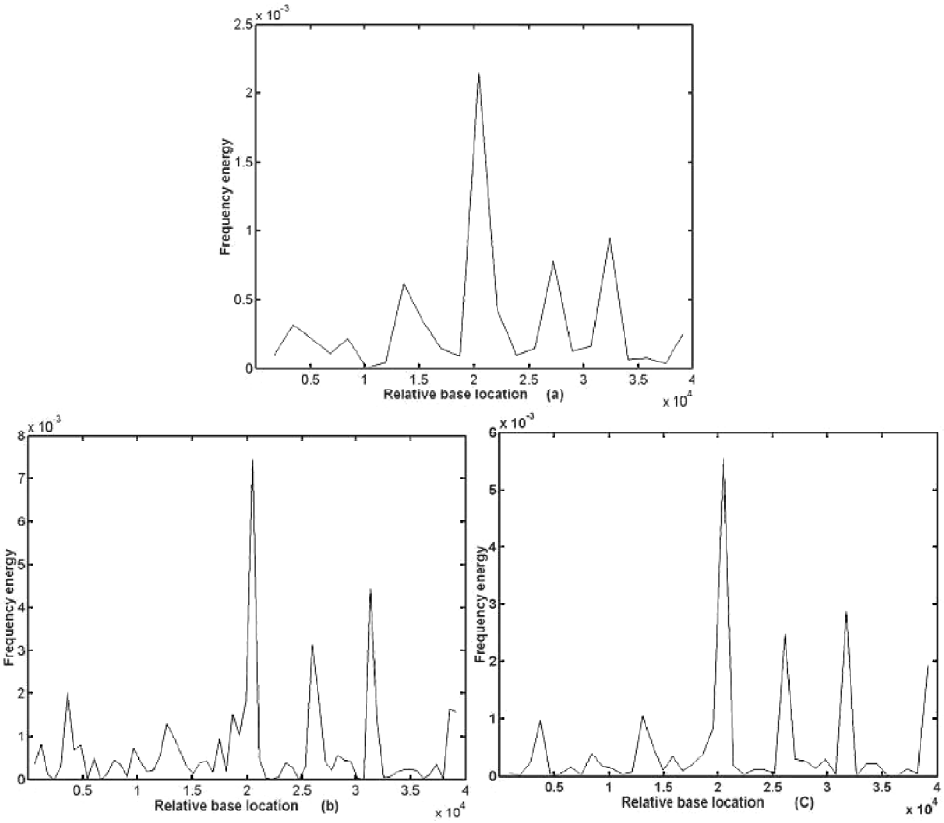


Fig. Frequency energy based detection algorithm (a subsequence of F56F11.4a with a length of 40,000 bases): (a) DFT. (b) WVD. (c) RSPWVD.

References

Anastassiou D. Frequency-domain analysis of Biomolecular sequences // BioInformatics. 2002. V. 16. P. 1073–1081.

Anthony G.J.F., Miller J.H., Suzuki D.T., Lewontin R.C., Gelbart W.M. Introduction to Genetic Analysis. New York: W.H. Freeman & Co., 1999.

Auger F., Flandrin P. Improving the readability of time-frequency and time-scale representations by the reassignment method, IEEE Transaction on Signal // Proc. 1995. V. 43, N 5. P. 1068–1089.

Cohen L. Time-Frequency Analysis, Prentice Hall, 1995.

Coward E. Equivalence of two Fourier methods for biological sequence // J. of Mathematical Biol. 1997. V. 36. P. 64–70.

John G.P., Dimitris G.M. Digital Signal Processing, Prentice Hall. 1996.

SEARCH FOR REGULATORY MOTIFS IN *DROSOPHILA MELANOGASTER* GENOME

***Samsonova A.*[1], *Dieterich C.*[2], *Vingron M.*[2], *Brazma A*[1].**

[1] EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK; [2] Max-Planck-Institute for Molecular Genetics, Ihnestraße 73, 14195 Berlin, Germany

***** Corresponding author: e-mail: nastja@ebi.ac.uk

Keywords: *Drosophila, development, gene expression, microarray, sequence motif search, SPEXS, twist and Dorsal regulatory cascades*

Summary

Motivation: The major challenge in understanding genomic sequences is to reveal how spatio-temporal information on gene expression is encoded in the genome. Identification of regions on the genome that contain regulatory information is the first stage of this process. It is widely accepted that comparative sequence data can help in location of these regions in higher eucaryotic genomes. The inter-species comparison allows us to identify highly conserved regions on the genome that can contain clusters of binding sites for multiple transcription factors.

Results: With published genomic sequences for *Drosophila melanogaster* and *Drosophila pseudoobscura*, we identified conserved regions on the genome that contain putative transcription factor binding sites. For the clusters of co-expressed genes found via microarray data analysis, we ran searches for statistically over-represented motifs in the corresponding groups of non-coding conserved sequences. These searches were carried out with SPEXS, which enumerates and scores all shared motifs in the sequences of unrestricted length.

Availability: Sequence motif data is available from the authors.

Introduction

The development of a multicellular organism is the succession of precisely orchestrated domain- and tissue-specific gene expression. It is known that spatio-temporal information on gene expression is encoded in the organism's genomic sequence. Although, many metazoan genomes are sequenced we are still far from understanding genome organization and transcriptional regulation. To understand development it is necessary to decipher the logic and organization of transcriptional networks. Identification of sequence motifs that call forth tissue-specific expression is a major challenge nowadays.

The modular organization of regulatory regions on the genome sequence has been reported by several research groups (e.g. Berman *et al.*, 2002). We decided to employ inter-species comparison to identify regions containing the unusually high concentration of transcription factor binding sites. It has been shown (Bergman *et al.*, 2002) that *D. pseudoobscura* is the best organism in genus *Drosophila* for comparative analysis and enchancer prediction. Recently sequenced *D. pseudoobscura* genome provides additional means to discriminate biologically relevant regions in the genome from those that occur by chance. We use highly conserved non-coding sequences found in an inter-species comparison for the search of transcription factor binding sites.

Formation of muscles during embryonic development of the fly is a complex process that requires synchronous action of many genes. The somatic muscles, the heart, the fat body, the somatic part of the gonad and most of the visceral muscles are derived from a series of segmentally repeated primordia in the *Drosophila* mesoderm. Multiple steps of fly mesoderm development are controlled by expression of bHLH transcription factor called *twist*, which is a regulatory target of a maternal

gene *Dorsal* and, in particular, is responsible for specification of muscle types. The role of *twist* in mesoderm development has been conserved during evolution. Moreover, some genes regulated by *twist* have highly conserved sequence and function in vertebrates (Furlong *et al.*, 2002).

Here, we demonstrate the applicability of the approach based on genome comparisons and exhaustive motif search to identify the transcription factor binding sites in *Dorsal* and *twist* gene regulatory cascades.

Materials and Methods

Full genomic sequence for *D. melanogaster* was downloaded from the Berkeley *Drosophila* Genome Project (BDGP) web site in FASTA format. We are currently using the BDGP Rel. 3.1 annotation and sequences. To determine the conserved regions on the *D. melanogaster* genome we use the scaffold of *D. pseudoobscura* genome Rel. 1.0 from Baylor College of Medicine.

The microarray gene expression data were obtained from Dr. E. Furlong's Lab, EMBL. This data set is a whole genome microarray describing gene expression during embryogenesis in the fly. To determine the conserved regions in *D. melanogaster* genome that can be used for finding "putative" regulatory regions we used the inter-species comparison data, obtained with the Waterman-Eggert algorithm. These alignments were obtained with two similarity rates. The first one is 95 % similarity, while the second one is 70 % similarity. We use the 70 % similarity rate sequences and their chromosome localization when searching for regulatory motifs, because the choice of the higher similarity rate results in the loss of some biologically important groups of genes.

All conserved sequences were re-annotated by means of BDGP Release 3.1. We have included into our analysis only those sequences that were annotated as non-coding regions.

The initial size of the raw conserved sequence data set is approximately 54.4Mb (50 % of the non-coding genome sequence); after annotation and filtration stages the new annotated conserved sequence data set contains only 37.5Mb. To reduce the level of noise only non-coding conserved sequences shorter than 2.5Kb were included in the analysis (see Fig.).

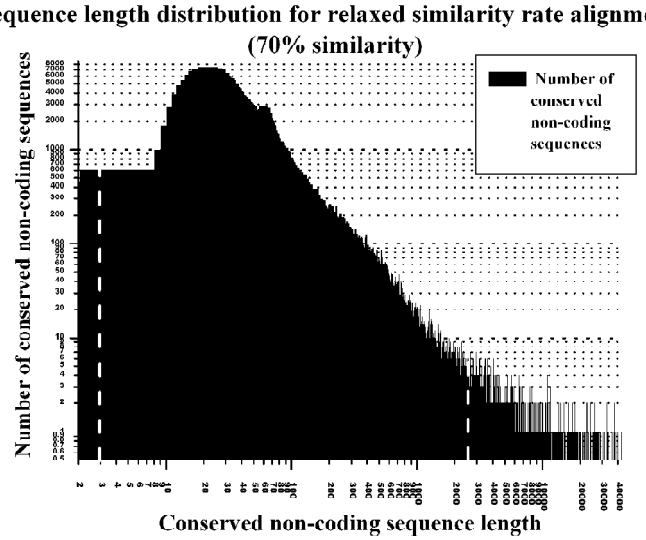


Fig. Non-coding sequence length distribution for conserved alignments. White dashed lines indicate the interval of lengths from 3 bp to 2500 bp for the sequences included in analysis and are eligible for motif search.

Our method is based on two assumptions: 1. We assume that co-expression is correlated with co-regulation; 2. If genes are co-regulated, then they should share clusters of common motifs.

We searched the non-coding conserved sequences of limited length for sequence motifs statistically over-represented within the sets of co-expressed genes from *twist* (Dr. E. Furlong, personal communication) and *Dorsal* (Stathopoulos *et al*., 2003) cascades.

Conserved non-coding sequences were extracted from the inter-species comparison data set. Searches were performed with SPEXS tool described in (Vilo, 2002). Given the set of conserved non-coding sequences this tool searches exhaustively for all possible sequence motifs, common to a minimum number of sequences in the set. For each motif found SPEXS calculates the statistical significance of its occurrence with respect to the control set of sequences. We take all non-coding conserved sequences as a control set.

The statistical significance was calculated according to the binomial distribution (Vilo *et al.*, 2000). To estimate significance thresholds, we repeated motif search on random data sets selected randomly from the control set and containing the same number of sequences as a test set. The randomization was repeated several times independently. We took into consideration motifs having binomial probabilities, ten times smaller than the lowest probability in the randomized set of the same size, and the ratio of over-representation higher than two.

Results and Discussion

Given a key role of *Dorsal* and *twist* in the mesoderm development we have chosen the regulatory cascades of these genes to search for tentative muscle-specific transcription factor binding sites. First, we used microarray data to identify clusters of co-regulated genes. For each gene present in the cascades we extracted a cluster of "the closest" genes in terms of expression profile, i.e. using Pearson centered distance measure we construct a distance matrix, which allows us to identify most similar or closest gene expression profiles. Second, for all genes in each of these clusters we constructed so-called sequence clusters, using corresponding conserved non-coding sequences from our data set. Finally, we perform a search for over-represented shared sequence motifs in these sequence clusters.

We have found both known transcription factor binding sites and novel motifs for genes involved in muscle development. Here, we show the best sequence motifs found in both cascades.

The search for over-represented motifs in clusters of co-expressed genes from the *Dorsal* cascade resulted in finding the GCTGTAGCT motif (binomial probability: 9.06592e-05, significance threshold: 0.000164011) in upstream sequences of genes *Scr*, *Abd-B* and *ex*; downstream sequences of genes *rho* and *Antp*; as well as *vn* and *nvy* introns. These genes are expressed in muscle tissue (*rho*, *Scr*, *Abd-B*, *Antp*, *vn*) or in nervous system (*nvy*, *Scr*, *Abd-B*, *Antp*) during stages 11–16 of development of the fly. Moreover, *rho*, *Scr*, *Abd-B* and *Antp* are homeobox genes and are reported to have common regulation. In both *Drosophila* and humans, homeobox clusters control the identity of cells along the anterior-posterior axis, including cells of ectodermal, mesodermal, endodermal and neural fates. It was also shown that *vn* interacts with *rho*.

When searching for over-represented motifs in the clusters of co-expressed genes from *twist* cascade we have found the sequence motif AGGCGTTAC (binomial probability: 1.02803e-07 and significance threshold: 0.00012939) in genes *eIF-4E*, *CG4022*, *CG1371*, *CG8806*. This is a well known muscle-specific consensus transcription factor binding site called *brinker.lax*. *Brinker* is an important gene expressed in mesoderm, and neuroectoderm during developmental stages 10–11. Genes *eIF-4E* and *CG1371* control the development of somatic and visceral muscles in the larva correspondingly, *CG4022* is expressed in central nervous system and *CG8806* is involved in mesoderm development.

To conclude, we have developed a pipeline that performs the search for putative transcription

factor binding sites in non-coding conserved sequences of co-expressed genes. The method is flexible enough and can be applied to more complex organisms than *D. melanogaster*. Results of the search described are biologically meaningful and are currently verified in experiments.

Acknowledgements

We thank Dr. E. Furlong from EMBL for providing us with the microarray gene expression data.

References

Bergman C.M. *et al*. Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome // Genome Biology. 2002. 3(12).

Berman B.P. *et al*. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome // PNAS. 2002. V. 99, N 2. P. 757–762.

Berkeley Drosophila Genome Project website. www.fruitfly.org.

Drosophila pseudoobscura genome. Human Genome Sequencing Center at Baylor College of Medicine website. ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Dpseudoobscura/fasta/.

Furlong E.E. *et al*. Gene expression during the life cycle of *Drosophila melanogaster* // Science. 2002. V. 298(5596). P. 1172.

Stathopoulos A., Levine M. *Dorsal* Gradient Networks in the *Drosophila* Embryo // Developmental Biology. 2003. V. 46(1). P. 57–67.

Vilo J. Pattern Discovery from Biosequences. PhD Thesis, Department of Computer Science, University of Helsinki, Finland Series of Publications A, Report A-2002-3 Helsinki. 2002.

Vilo J., Brazma A., Jonassen I., Robinson A., Ukkonen E. Mining for putative regulatory elements in the yeast genome using gene expression data // ISMB. 2000 August. AAAI Press, 2000. P. 384–394.

LONG SEGMENTAL REPEATS IN HUMAN GENOME: DENSITY, DISTRIBUTION, STRUCTURE

***Saraev D.V.*[1], Dzhekshenbaeva G.K.[1], Baksheyev D.G.[2], Rodionov K.V.[2], Golubitskii A.A.[2], Fursov M.Yu.[2], Golosov I.S.[2], Kisselev L.L.[3], Blinov V.M.*[1]**

[1] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia; [2] Novosibirsk Center of Information Technologies "UniPro", Novosibirsk, Russia; [3] Engelhardt Institute of Molecular Biology, Moscow, Russia

* Corresponding authors: e-mail: dsaraev@newmail.ru; vblinov@online.nsk.su

Keywords: *human genome, segmental repeats, translocations, recombination, distribution density, conservation/compensation laws*

Resume

Motivation: The search and study long segmental repeats is of basic meaning for evolutional and functional analysis of the human genome. *Results:* A new computer technology is developed for identification and analysis of long segmental repeats (LSR). The technology is essentially based on the compression of nucleic sequence data and allows us to represent whole human genome in 1000 times less space. With this technology we discovered 1098 LSRs in the human genome with lengths between 10k and 3.8M bp and similarity degree of 85 % to 99 %. The total length of the repeats comprised 4.40 % of the whole genome. Analysis of the distribution density and structure of the LSRs in each chromosome allowed us to classify the chromosomes with respect to the LSR copy number. Differentiating direct and inverted LSRs, as well as intra- and inter-chromosomal LSRs makes us conclude that the numbers of direct and inverted LSRs are equal in each chromosome and intra/interchromosomal LSRs are clustered around the same "hot spots," potentially the recombination and translocation sites.

Availability: none

Introduction

The study of direct and inverted long (>10k bp) segmental repeats is of fundamental importance for functional and structural analysis of the human genome. This study supports comparative evolutionary analyses of mammalian genomes and development of the interchromosomal translocation and recombination models (Blinov *et al*., 1998; Blinov *et al*., 2001).

According to some estimates, cumulative length of the LSRs is between 3 % and 5 % of the size of human genome (Eichler, 1998; Bailey *et al*., 2001; Cheung *et al*., 2003). To clarify the role of the LSRs an analysis of the distribution density and structure of these genetic elements has been increasingly demanded at both whole genome and chromosomal scales. Most up-to-date methods for discovering LSRs *in silico* have been based on BLAST (Altschul *et al*., 1990). In this report we present a computer technology for identification and analysis of long segmental repeats which is built upon the following principles: 1. Direct whole genome comparison based on DPview technology (Fursov *et al*., 2004); 2. *Ad hoc* compression of nucleic sequences data and superfast comparison of whole genomes encoded in a new 26-character alphabet (A, B,…, Z); 3. Classification of the LSRs into direct/inverted and intra/inter-chromosomal ones, which allows of large scale calculations of distribution density of these elements and establishing laws of conservation, compensation as well as interaction mechanisms between them; 4. Dot-matrix representation and visual interpretation of the comparisons (two-dimensional dot plot of the whole human genome displays all LSRs in all 24 chromosomes.

Methods and Algorithms

We studied NCBI builds 33 and 34 of the human genome. Genomic data encoding and compression was performed with an *ad hoc* software rpt.exe. Inexact homology dot-matrices were computed on a parallel supercomputer with MegaGene software (Blinov, *et al*., 1999); dot-matrices for exact matches were computed with DPview algorithm (Fursov *et al*., 2004). Alignments of the LSRs were carried out with help of Alignment Service (AS) software (Resenchuk *et al*., 1995). Distribution density, structure, and graphical representation of the LSRs in all 24 chromosomes, statistics, correlation analysis, and classification of the LSRs into homology groups were also obtained with the help of AS.

Results and Discussion

Analysis of the 33rd build of the human genome revealed 1098 LSRs with lengths from 10k to 3.8M bp with total length equal to 4.40 % of the size of whole genome. The most interesting observation is the dependence of the LSR length interval on their distribution density: the density decreases as the length of the LSRs increases (Fig.). Most represented in the density are LSRs with lengths in the range 100k to 400k bp, then the LSRs with lengths 10k 100k bp, whereas densities of 400k 1M and more than 1M bp repeats is practically the same (Fig.).

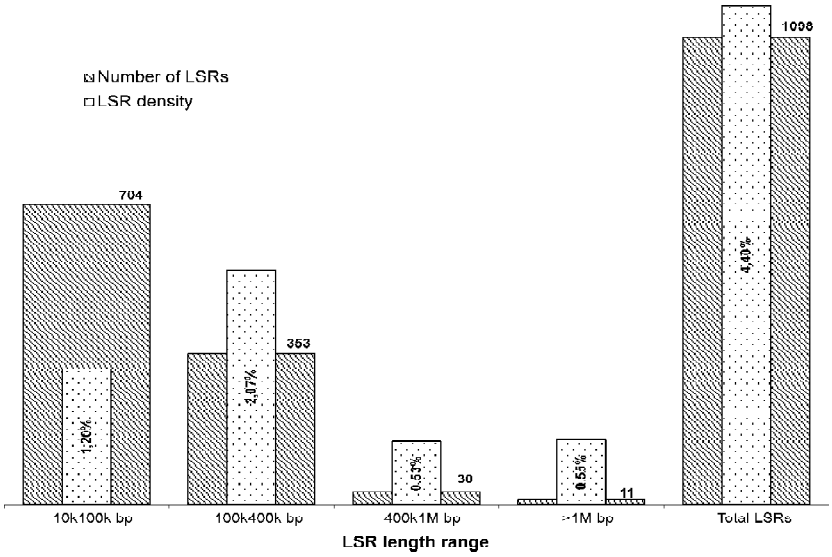


Fig. Distribution density and number of LSRs in the human genome.

With respect to LSR distribution density the chromosomes can be ordered as follows: 0Y-16-15-22-07-09-21-11-0X-08-04-02-17-19-20-10-01-18-05-13-12-03-06-14. In this sequence chromosomes 0Y-16-15-22-… have maximum content of LSRs and, contrarily, LSR distribution density is minimal in chromosomes 14-06-03-12-….

In the human genome LSRs are represented in direct(+) and inverted(–) orientations. Also LSRs can be divided to "intrachromosomal" and "interchromosomal." The analysis of the four types of LSRs for each of 24 chromosomes revealed the following properties. Distribution density in direct and inverted orientations along the chromosomes is almost identical: most of the (+) and (–) maxima reproduce each other, whilst intra- and interchromosomal LSRs are localized near the same "hot spots" in the genome.

These observations, not reported before, indicate of existence of "conservation and compensation laws" for LSRs in the human genome. We suggest that recombination processes control the distribution of LSRs. It is yet to be explained which genetic elements are mapped in these "hot spots" and what role they play in the emergence of long segmental repeats.

Table. Number and density of LSRs in human chromosomes

Chromosome	Number of LSRs	Total LSR length, bp	Chromosome length, bp	Length ratio, %
01	106	9 332 892	245 203 898	3.81
02	58	7 597 534	243 315 028	3.12
03	39	3 431 101	199 411 731	1.72
04	76	8 010 570	191 610 523	4.18
05	71	7 072 010	180 967 295	3.91
06	28	2 614 256	170 740 541	1.53
07	96	10 251 650	158 431 299	6.47
08	56	7 598 381	145 908 738	5.21
09	57	7 127 710	134 505 819	5.30
10	39	3 570 969	135 480 874	2.64
11	49	7 844 952	134 978 784	5.81
12	22	2 614 304	133 464 434	1.96
13	22	1 458 538	114 151 656	1.28
14	5	316 858	105 311 216	0.30
15	52	6 172 471	100 114 055	6.17
16	69	10 171 539	89 995 999	11.30
17	21	2 023 994	81 691 216	2.48
18	23	3 785 119	77 753 510	4.87
19	41	2 698 691	63 790 860	4.23
20	31	1 778 928	63 644 868	2.80
21	24	4 698 231	46 976 537	10.00
22	29	2 592 235	49 476 972	5.24
0X	23	6 232 772	152 634 166	4.08
0Y	61	16 134 063	50 961 097	31.66
Whole genome	1098	135 129 768	3 070 521 116	4.40

Acknowledgements

The work was supported by Russian Ministry of Science and Technology grant "Development of software for analysis of structural and functional properties of the human genome."

References

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. P. 403–410.

Bailey J.A., Yavor A.M., Massa H.F., Trask B.J., Eichler E.E. Segmental duplications: organization and impact within the current human genome project assembly // Genome Res. 2001. V. 11. P. 1005–1017.

Blinov V.M., Resenchuk S.M., Uvarov D.L., Chirikova G.B., Denisov S.I., Kisselev L.L. Alu elements in human genome: invariant secondary structure of left and right monomers // Mol. Biol. 1998. V. 32(1). P. 70–77 (in Russ.).

Blinov V.M., Resenchuk S.M., Denisov S.I., Chirikova G.B., Uvarov D.L., McCor-k-le S., Anderson C. MegaGene: a New Computer Technology for Analyzing Complete Viral Genomes // XIth International Congress of Virology, Sydney, VW63.02, 1999. P. 152.

Blinov V.M., Denisov S.I., Saraev D.V., Shvetsov D.V., Uvarov D.L., Opa-ri-na N.Yu., Sandakhchiev L.S., Kisselev L.L. Structural organization of human genome: distribution of nucleotides, Alu repeats, and exons in chromosomes 21 and 22 // Mol. Biol. 2001. V. 35(6). P. 1–7 (In Russ.).

Cheung J., Estivill X., Khaja R., MacDonald j.R., Lau K., Tsui L.-C., Scherer S.W. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence // Genome Biol. 2003. V. 4.

Eichler E.E. Masquerading repeats: paralogous pitfalls of the human genome // Genome Res. 1998. V. 8(8). P. 758–762.

Fursov M.Yu., Baksheyev D.G., Rodionov K.V., Golubitskii A.A., Saraev D.V., De-ni-sov S.I., Blinov V.M. A practical method for maximum exact matches in large genomes // Submitted to 4th International Conference on Bioinformatics of Genome Regulation and Structure (BGRS 2004). 2004.

Resenchuk S.M., Blinov V.M. Alignment service: creation and processing of alignments of sequences of unlimited length // CABIOS. 1995. V. 11(1). P. 7–11.

GENOME-SCALE PREDICTION OF TRANSCRIPTION FACTORS AND THEIR TARGETS

***Sarai A.*[1], *Ahmad Sh.*[1], *Gromiha M.M.*[2], *Kono H.*[3]**

[1] Dept. Bioscience & Bioinformatics, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, 820-8502 Japan; [2]Computational Biology Research Center, AIST, 2-41-6 Koto-ku, Tokyo 135-0064 Japan; [3]Neutron Science Research Center and Center for Promotion of Computational Science and Engineering Japan Atomic Energy Research Institute, 8-1, Umemidai, Kizu-cho, Soraku, Kyoto, 619-0215 Japan
* Corresponding author: e-mail: sarai@bse.kyutech.ac.jp

Keywords: *transcription factor, target genes, genome*

Summary

Motivation: Complete genome sequences of many organisms have become available and the functional analysis of genomes is a target of intensive research. Gene regulation in higher organisms is one of the most important biological functions. Identifying transcription factors and finding their target genes at the genome level will lay a basis for the analysis of the gene regulatory network.

Results: We have developed an algorithm for predicting DNA-binding proteins based on sequence and structural information. We have achieved accuracies higher than 80 %. For the prediction of target sites, we have used a knowledge-based approach, utilizing rapidly increasing structural data of protein-DNA complexes to derive empirical potential functions for the specific interactions between bases and amino acids as well as for DNA conformation, by the statistical analyses of the structural data. These statistical potentials were used to quantify the specificity of protein-DNA recognition, which enabled us to establish the structure-function relationship of transcription factors, such as the effects of binding cooperativity on target recognition. The method was applied to yeast genome sequences, and we could identify target genes of transcription factors successfully. We are also developing an integrated genome-scale prediction system by combining various kinds of methods.

Availability: The prediction system will be made available to the public soon.

Introduction

The genome analyses show that in most species about a half of the genes is of function unknown. Many of the genes may turn out to code for transcription factors. Subsequent functional analyses of transcription factors involve identification of their target genes. Transcription factors usually bind to multiple target sequences and regulate multiple genes in a complex manner. However, the targets of transcription factors are largely unknown. Thus, understanding the molecular mechanism of protein-DNA recognition and its application to genome-wide prediction are essential for the analysis of gene regulation network. We have been developing methods for predicting transcription factors and their targets based on various kinds of information ranging from sequence to structure. Here we present a strategy for genome-scale prediction of transcription factors and their targets based a combination of these methods.

Methods

In order to predict whether the new protein products derived from genome analysis bind to DNA or not if they do not have known homologues, we have developed a prediction method based on sequence and structural information (Ahmad *et al*., 2003; Ahmad, Sarai, 2004). For the prediction of targets of transcription factors, we made a statistical analysis of structural database of protein-

DNA complex, and derived empirical potential functions for the specific interactions between bases and amino acids (Kono, Sarai, 1999; Sarai *et al*., 2001; Kono, Sarai, 2003; Selvaraj *et al*., 2002). Then, we used a sequence-structure threading to examine the relationship between structure and specificity in protein-DNA recognition. By threading a set of random DNA sequences onto the template structure, we calculated the Z-score of the specific sequences against the random sequences, which represent the specificity of the complex. By threading real genome sequences, we can predict target sites of transcription factors at the genome scale. In addition to this so-called direct or "intermolecular" readout mechanism, by which proteins recognize DNA sequence through the direct contact between amino acids and base pairs, we also evaluated the fitness of DNA sequence against DNA structure to examine the role of so-called indirect or "intramolecular" readout mechanism (hereafter we use the terms "intermolecular" and "intramolecular" readout mechanisms). For this purpose, we have derived statistical potential functions for conformational energy of DNA from the same set of protein-DNA complex structures (Sarai *et al*., 2001; Gromiha *et al*., 2004). In another approach, we have analyzed protein-DNA recognition by computer simulations. We have used various kinds of methods: Monte Carlo simulation of base amino acid interactions (Pichierri *et al*., 1999; Sayano *et al*., 2000; Yoshida *et al*., 2002); molecular dynamics of DNA conformation; empirical calculations of interaction energy of protein-DNA complex (Oobatake *et al*., 2003): molecular dynamics/free energy calculations of protein-DNA complex (Saito, Sarai, 2003); and docking simulation of protein-DNA binding and sliding.

Results and Discussion

In order to predict DNA-binding proteins, we have used amino acid composition, sequence information, solvent accessibility surface information (Ahmad *et al*., 2003) and electrostatic information such as charge and dipole moment (Ahmad, Sarai, 2004). We have achieved accuracies higher than 80 %. These simple and yet accurate methods, together with homology modelling of structures, enable us to predict DNA-binding proteins at the genome scale quickly.

We have derived empirical potential functions for the base-amino acid interactions from the analysis of protein-DNA complex structures. We have compared the structures of cognate and non-cognate protein-DNA complex structures in order to test our method and to understand what is important for specific binding and what is different between them. The statistical potentials could distinguish the two structures as differences in the Z-scores as well as statistical potentials (Kono, Sarai, 1999; Selvaraj *et al*., 2002). Thus, the subtle differences in specificity of these structures could be detected by our method. We also applied this method to examine the relationship between structure and specificity in cooperative protein-DNA binding. The effect of cooperative binding was examined by comparing the monomer and heterodimer complexes of MATa1/α2 (Kono, Sarai, 1999), MCM1/MATα2 and NFAT/AP-1 transcription factors. We found that the heterodimer binding enhances the specificity in a non-additive manner. This result indicates that the conformational changes introduced by the heterodimer binding play an important role in enhancing the specificity.

We can calculate Z-score for the intramolecular recognition due to sequence-dependent DNA conformation in the same way as for the intermolecular recognition. By comparing both the Z-scores we can assess the relative contributions of intermolecular and intramolecular readout mechanisms. We have analyzed various protein-DNA complexes systematically, and found that both the intermolecular and intramolecular mechanisms make significant contribution to the specificity (Sarai *et al*., 2001; Gromiha *et al*., 2004). The relative contributions depend on the types of DNA-binding proteins. Because both the potentials are independent quantities, they can be summed up to calculate the total energy and used to find target sites, although a weighting factor needs to be determined as the two potentials were derived from different statistics. We found that the Z-score was indeed enhanced compared with individual Z-scores for intermolecular

or intramolecular readout alone. This result indicates that the energies of the intermolecular and intramolecular readouts contain independent information that in combination enhances the specificity of the recognition. We used the combined energy for threading against DNA sequences to make target predictions for transcription factors.

The threading procedure was used to find target sites of transcription factors in real genome sequences. As an example of such applications, we could identify the experimentally-verified binding sites of the transcription factor MATa1/α2 in the promoter of HO gene successfully (Kono, Sarai, 1999). We have also attempted to identify target sites and genes of MCM1/MATα2 in the whole yeast genome. The target genes of this transcription factor have been identified in yeast genome experimentally (Zhong, Vershon, 1997). The predicted target genes were ranked by the Z-score and compared with experimental data. The target genes identified positively by experiment were ranked high in the list, and the experimentally negative genes were ranked low (Sarai *et al*., 2004). Separation between the positive and negative genes was not perfect but they were segregated by a certain threshold Z-score value. The total Z-score gave better separation than that of direct contribution alone.

In order to complement the knowledge-based approach described above, we also performed various kinds of computer simulations (Pichierri *et al*., 1999; Sayano *et al*., 2000; Yoshida *et al*., 2002; Oobatake *et al*., 2003; Saito, Sarai, 2003). These simulations could derive interaction potentials between bases and amino acids equivalent to the statistical potentials, and provided insight into the molecular mechanism of specificity in protein-DNA recognition. We are now combining knowledge-based approach and computer simulations, with other methods based on sequence information and experimental binding data (Deng *et al*., 1996) and genome annotation information, to develop an automated prediction system. Such integrated prediction system with combination of different kinds of methods would become a powerful tool for analyzing transcription factors and for providing insight into the mechanism of gene expression regulation.

References

Ahmad S., Gromiha M.M., Sarai A. Role of composition, sequence and structural information in DNA-binding: analysis and prediction // Bioinformatics. 2003. V. 19. P. 1849–1851.

Ahmad S., Sarai A. Moment based prediction of DNA-binding proteins // J. Mol. Biol. 2004, to be published.

Deng Q.-L., Ishii S., Sarai A. Binding-site analysis of c-Myb: screening of potential binding sites by the mutational matrix derived from systematic binding affinity measurements // Nucleic Acids Res. 1996. V. 24. P. 766–774.

Gromiha M.M., Siebers J.G., Selvaraj S., Kono H., Sarai A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition // J. Mol. Biol. 2004. V. 337. P. 285–294.

Kono H., Sarai A. Structure-based prediction of DNA target sites by regulatory proteins // Proteins. 1999. V. 35. P. 114–131.

Kono H., Sarai A. Structure-specificity relationship in protein-DNA recognition // Bioinformatics of Genome Regulation and Structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic, NLD, 2003.

Oobatake M., Kono H., Wang Y-F., Sarai A. Anatomy of specific interactions between lambda repressor and operator DNA by energy-component analysis // Proteins. 2003. V. 53. P. 33–43.

Pichierri F., Aida M., Gromiha M.M., Sarai A. Free energy maps of base-amino acid interaction for protein-DNA recognition // J. Am. Chem. Soc. 1999. V. 121. P. 6152–6157.

Saito M., Sarai A. Free energy calculations for the relative binding affinity between DNA and lambda repressor // Proteins. 2003. V. 52. P. 129–136.

Sarai A., Selvaraj S., Gromiha M.M., Siebers J.-G., Prabakan P., Kono H. Target prediction of transcription factors: refinement of structure-based method // Genome Informatics. 2001. V. 12. P. 384–385.

Sarai A., Siebers J., Selvaraj S., Gromiha M.M., Kono H. Integration of bioinformatics and computational biology to understand protein-DNA recognition mechanism // J. Bioinformatics and Computational Biology. 2004, to be published.

Sayano K., Kono H., Gromiha M.M., Sarai A. Multicanonical Monte Carlo calculation of free-energy map for base-amino acid interaction // J. Compt. Chem. 2000. V. 21. P. 954–962.

Selvaraj S., Kono H., Sarai A. Specificity of protein-DNA recognition revealed by structure-based potentials: symmetric/asymmetric and cognate/noncognate binding // J. Mol. Biol. 2002. V. 322. P. 907–915.

Yoshida T., Nishimura T., Aida M., Pichierri F., Gromiha M.M., Sarai A. Evaluation of free energy landscape for base-amino acid interactions using ab initio force field and extensive sampling // Biopolymers. 2002. V. 61. P. 84–95.

Zhong H., Vershon A.K. The yeast homeodomain protein MATalpha2 shows extended DNA binding specificity in complex with Mcm1 // J. Biol. Chem. 1997. V. 272. P. 8402–8409.

BINARY TREE FOR CLUSTERING OF REGULATORY SIGNALS

***Stavrovskaya E.D.*[1], *Mironov A.A.*[2, 3]**

[1] Department of Information Technologies, Peoples' Friendship University of Russia, 117198, Moscow, Russia; [2] State Scientific Center GosNIIGenetika, Moscow, 113545, Russia; [3] Department of Bioengineering and Bioinformatics, Moscow State University, 119992, Moscow, Russia
* Corresponding author: e-mail: esta191@fromru.com

Keywords: *regulation signal, regulon, cluster*

Resume

Motivation: Application of phylogenetic footprinting techniques to the analysis of the bacterial regulatory signals results in generation of sets of conserved candidate sites found upstream of orthologous genes. The next step of such analysis should be clustering of sites corresponding to the same signal and thus likely being the binding sites for the same regulator.

Results: Here we present an algorithm for clustering of candidate regulatory sites leading to identification of groups of co-regulated genes.

Availability: Algorithm was realized using C++ language as console application for Win32. Source code is available upon request.

Introduction

Analysis of regulatory systems is the one of the actual problems of modern molecular biology. The first step in such analysis is identification of co-regulated genes. Generally, binding sites of one regulator are similar. Thus, by clustering regulatory sites, it is possible to reveal potential regulons.

The existing methods of sequence motif clustering can be subdivided into hierarchical (Pietrokovski, 1996; Hughes *et al*., 2000) and methods based on statistical models (van Nimwegen *et al*., 2002; Qin *et al*., 2003). Here we present an algorithm using binary tree for clustering regulatory sites. This method is hierarchical and uses the same tree derivation strategy as in (Hughes *et al*., 2000), but unlike the latter it uses an original procedure to detect the clusters on the tree.

Algorithm

The algorithm consists of two steps: tree derivation and tree-walk.

Each tree node is associated with a set of sites and each leaf of the tree corresponds to one of constructed sites. The tree derivation is an iterative process. At each iteration a set of subtrees is considered. The distance between the site sets is computed as the correlation between the positional nucleotide frequency matrices.

A pair of the closest subtrees is merged, forming one subtree. This results in a binary tree whose root corresponds to the set of all candidate sites.

At the tree walk stage, the information content (Gelfand *et al*., 2000) of each node (set of sites) is calculated as

$$I = \sum_{k=1}^{l} \sum_{i=A,C,G,T} f(i,k) \log[f(i,k)/0.25] \qquad f(i,k) = (n(i,k) + 0.25\sqrt{N})/(N + \sqrt{N}),$$

where *f(i, k)* is the relative frequency of nucleotide *i* at position *k*, $0.25\sqrt{N}$ and $\sqrt{N}$ are the pseudocounts.

We define cluster (signal) as a node that corresponds to a local maximum of the information content so that the node satisfies the condition:

$$(I > I_p) \& ((I > I_l) \vee (I > I_r)),$$

where I is the information content of a node, I_p is the information content of the parent node, I_l and I_r are the information contents of its left and right child nodes.

Pseudocounts are introduced in order to avoid the trivial solution (each leaf is a cluster).

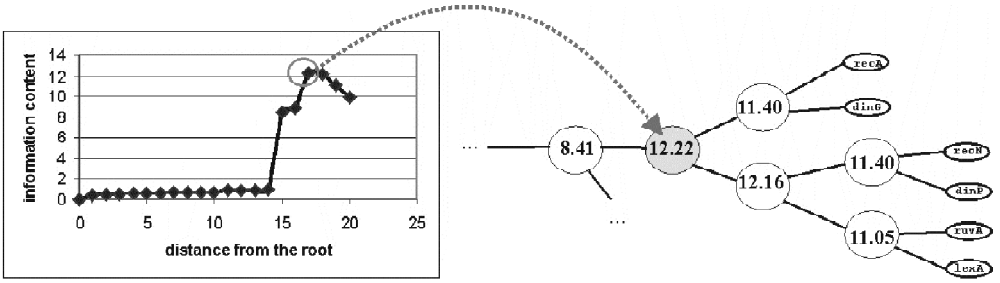


Fig. The dependence of the information content on the distance from the root.

The Figure shows the dependence of the information content on the distance from the root. The tree node after the abrupt change corresponds to a cluster.

Table 1.

Regulator	Gene	Pos.	Site sequence	Number of known sites in sample	Number of sites in a found cluster	Number of known sites in a cluster
lexA Inf : 12.98	EC_umuD	161	ctactgtatataaaaacagtat	7	7	7
	EC_recN	67	ttactgtatataaaaccagttt			
	EC_lexA	112	ttgctgtatatactcacagcat			
	EC_dinP	86	tcactgtatactttaccagtgt			
	EC_ruvA	131	tcgctggatatctatccagcat			
	EC_recA	50	atactgtatgagcatacagtat			
	EC_ding	135	atattggctgtttatacagtat			
purR Inf : 13.28	EC_purM	119	gtctcgcaaacgtttgctttcc	6	6	6
	EC_purH	75	gttgcgcaaacgttttcgttac			
	EC_purE	78	gccacgcaaccgttttccttgc			
	EC_cvpA	126	cctacgcaaacgttttcttttt			
	EC_purR	138	taaaggcaaacgtttaccttgc			
	EC_purL	106	tccacgcaaacggtttcgtcag			

Testing and Results

The algorithm was applied to upstream regions of orthologous genes from gamma proteobacteria.

Table 2.

Regulator	Number of known sites in the sample	Number of sites in a cluster	Number of known sites in a cluster
cpxR	**12**	**10**	**10**
crp	**49**	**78**	**49**
fadR	**7**	**6**	**6**
flhCD	**3**	**3**	**3**
fur	**9**	**6**	**6**
gcvA	**4**	**4**	**4**
hipB	**4**	**3**	**3**
lacI	**3**	**3**	**3**
lexA	**19**	**15**	**15**
metJ	**15**	**6**	**6**
metR	**8**	**8**	**8**
modE	**3**	**3**	**3**
nagC	**6**	**4**	**4**
ntrC	**5**	**5**	**5**
phoB	**15**	**12**	**12**
purR	**22**	**16**	**16**
trpR	**4**	**4**	**3**
tyrR	**17**	**9**	**9**

Sets of putative regulatory sites were constructed by (Danilova *et al*., 2001). There were sites of length 22 only. Sample results are given in Table 1. These clusters contain all sites known to be related to a regulator from the sample and no alien sites.

We also tested the algorithm on regulatory sites from database DPInteract (Robison *et al*., 1998). This sample contains sites of different lengths. The results are listed in Table 2.

Table 1 and Table 2 demonstrate that the algorithm gives reasonable results in general. However, clusters for some regulators are not complete. It is normal for weak signals like crp and tyrR. But it is surprising in cases of strong signals like purR. The reason for that is that cluster of higher level contains also sites for regulator galR. They are mixed with purR sites and thus exclude some sites from the purR cluster. This is due to difficulties regarding clustering of sites of different lengths.

Clustering sites of different lengths

Clustering of sites having different lengths presents some problems. It is necessary to compare sets of sites of different lengths. Moreover, sets of longer sites have a large information content, and thus a local maximum of the information content may not correspond to a cluster. The solution of the first problem is simply to supplement shorter sites with flanking positions in the genome. To solve the second problem we propose to compute the statistical significance of positions in the set of sites and sum the information content of significant positions only.

Statistical significance

The information content of a position in a set of sites depends on the distribution of nucleotides in this position. The probability to observe distribution (k_1, k_2, k_3, k_4) is

$$f(\inf = I) = f(k_1, k_2, k_3, k_4) = \frac{C_N^{k_1} 4 C_{N-k_1}^{k2} u_3 C_{N-k_1-k_2}^{k3} u_2}{4^N Q(k_1, k_2, k_3, k_4)}$$

or

$$f(k_1, k_2, k_3, k_4) = \frac{N! u_3 u_2}{k_1! k_2! k_3! 4^{N-1} Q(k_1, k_2, k_3, k_4)},$$

where

$$u_3 = \begin{cases} 3, k_2 > 0 \\ 1, k_2 = 0 \end{cases}, \; u_2 = \begin{cases} 2, k_3 > 0 \\ 1, k_3 = 0 \end{cases}, \; Q(k_1, k_2, k_3, k_4) = \begin{cases} 4!, & k_1 = k_2 = k_3 = k_4 \\ 3!, & k_1 = k_2 = k_3 > k_4, or \\ & k_1 > k_2 = k_3 = k_4 > 0 \\ 2!, & k_1 = k_2 > k_3 > k_4, or \\ & k_1 > k_2 = k_3 > k_4, or \\ & k_1 > k_2 > k_3 = k_4 > 0 \\ 2!2!, & k_1 = k_2 > k_3 = k_4 > 0 \\ 1, & in\ all\ other\ cases \end{cases}$$

where k_i is the number of nucleotides of type i, $k_1 \geq k_2 \geq k_3 \geq k_4$, $k_1+k_2+k_3+k_4=N$ is the number of sites in the set (column height).

For selection of significant positions we use the Bernoulli estimator (Kalinina *et al*., 2004) to set the threshold value. Order the observed information contents by decrease: I_1, I_2, K , and find k^* such that:

$$k^* = \arg\min_k P(\text{there are at least } k \text{ observed information contents } I \geq I_k) = \arg\min_k \left(1 - \sum_{i=n-k+1}^{n} C_n^i q^i p^{n-i}\right),$$

where n is the total number of considered positions, $p = P(I \geq I_k) = \sum_{I \geq I_k} f(\inf = I)$,

$q = 1 - p$.

The k^* is the desired number of significant positions.

Acknowledgements

We are grateful to L.Danilova for data and to M.Gelfand and D.Ravcheev for useful discussions. This study was partially supported by grants from HHMI (55000309) and the Program in Molecular and Cellular Biology (RAS).

References

Danilova L.V., Gorbunov K.Yu., Gelfand M.S., Lyubetskii V.A. Algorithm of regulatory signal recognition in DNA sequences // Mol. Biol. 2001. V. 35. P. 987–995.

Gelfand M.S., Koonin E.V., Mironov A.A. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach // Nucleic Acids Res. 2000. V. 28. P. 695–705.

Hughes J.D., Estep P.W., Tavazoie S., Church G.M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae // J. Mol. Biol. 2000. V. 296(5). P. 1205–1214.

Kalinina O. V., Mironov A. A., Gelfand M.S., Rakhmaninova A.B. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families // Protein Science. 2004. V. 13(2). P. 443–56.

Nimwegen E. van , Zavolan M., Rajewsky N., Siggia E.D. Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics // Proc. Natl Acad. Sci. USA. 2002. V. 99(11). P. 7323–7328.

Pietrokovski S. Searching databases of conserved sequence regions by aligning protein multiple-alignments // Nucleic Acids Res. 1996. V. 24(19). P. 3836–3845.

Qin Z.S., McCue L.A., Thompson W., Mayerhofer L., Lawrence C.E., Liu J.S. Identification of co-regulated genes through Bayesian clustering of predicted regulatory binding sites // Nat. Biotechnol. 2003. V. 21(4). P. 435–439.

Robison K., McGuire A.M., Church G.M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome // J. Mol. Biol. 1998. V. 284(2). P. 241–254.

THEORETICAL AND EXPERIMENTAL STUDY OF MUTATIONS INDUCED BY 8-OXOGUANINE

Vasyunina E.A.*[1]*, *Rogozin I.B.*[1,2], *Sinitsina O.I.*[1], *Plaksina A.S.*[1], *Rotskaya U.N.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [2] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
* Corresponding author: e-mail: evas@niboch.nsc.ru

Keywords: *oxidative mutagenesis, evolutionary conservation, DNA context, hotspots*

Summary

Motivation: The process of mutation is fundamental in biology, being an essential evolutionary factor that creates genetic variation. Understanding complex mechanisms by which mutations occur spontaneously or are induced by mutagens is an important goal of molecular biology.

Results: We analyzed mutations induced by 8-oxoguanine (8-oxoG) using various computational techniques. Results suggested that mutagenesis due to 8-oxoG is significantly influenced by nearest neighboring bases and the context is quite evolutionarily stable. The revealed context properties might reflect intrinsic properties of interactions between 8-oxoG and DNA. Comparative analysis of spontaneous mutations suggested that a substantial fraction of spontaneous A•T→C•T mutations is caused by 8-oxoGTP in nucleotide pools.

Introduction

Mutation frequencies vary significantly along nucleotide sequences such that mutations often concentrate at certain positions called hotspots. Mutation hotspots in DNA reflect intrinsic properties of the mutation process, such as sequence specificity, that manifests itself at the level of interaction between mutagens, DNA, and the action of the repair and replication machineries (Drake, Baltz, 1976; Rogozin, Pavlov, 2003). In this study, we analyzed mutations induced by 8-oxoguanine (8-oxoG). Chemical agents, ionizing radiation and oxidative stress cause DNA oxidation (Adelman *et al*., 1988). 8-OxoG is one of the most prominent base oxidation products and has been implicated in mutagenesis, carcinogenesis and aging B.N. (Ames, 1989). It has been shown to cause G•C→T•A and A•T→C•G mutations *in vivo* and *in vitro*, depending whether guanine is oxidized in DNA or in the DNA precursor pools, respectively (Michaels *et al*.,. 1992; Pavlov et al., 1994). To counter the mutagenic effects of 8-oxoG, *E. coli* has an effective repair system containing three genes, *mutT*, *mutM* and *mutY* (Michaels, Miller, 1992). *mutM* and *mutY* are involved in repair of 8-oxoG in DNA and *mutT* codes for an enzyme that converts 8-oxoGTP in the nucleotide pool to 8-oxoGMP, preventing the incorporation of 8-oxoG into DNA (Maki, Sekiguchi, 1992).

Results and Discussion

A spontaneous mutation spectrum in the *mutT* deficient *E. coli* strain (*the ECDEFT entry from the DBMS database, ftp.bionet.nsc.ru/pub/biology/dbms/*) is composed almost exclusively, of A•T→C•G transversions which is in general consistent with mutagenic properties of 8-oxoGTP (Fowler and Schaaper, 1997; Tassotto and Mathews, 2002). Hotspot context analysis of these transversions (Fowler, Schaaper, 1997) using regression trees (Berikov, Rogozin, 1999) revealed AA mutable sequence (the hotspot position is underlined) (Fig.).

Comparison of the *mutT* spectrum and A•T→C•T transversions in a spectrum of spontaneous mutations in the *lacI* gene (*lacI*^{q} test system) (Oller, Schaaper, 1994; *the ECLACSP entry, ftp.bionet.nsc.ru/pub/biology/dbms/*) did not reveal significant differences between them (Table 1). Furthermore, a highly significant positive correlation was found (Table 1). This result suggested that a substantial fraction of spontaneous A•T→C•T mutations in *E. coli* is caused by 8-oxoGTP in nucleotide pools.

Position	Site sequence	Number of mutations	Position	Site sequence	Number of mutations
	- - * - -				
77	A T a A G	20	81	T C a G A	10
167	G T a A T	37	83	A G a C C	5
189	A C a A C	23	87	A A a C G	4
192	A C a A C	18	96	C C a C G	5
	---------		105	C C a G G	2
	N N a A N		110	C C a G C	0
	Hotspots sites		117	A A a C G	1
			128	A A a C G	4
	- - * - -		141	C C a C T	5
41	T A a C G	4	168	T T a C A	9
54	C G a C A	8	177	C A a C C	7
64	G C a T A	2	190	C A a C A	0
72	A G a C A	5	195	C C a G T	10
79	T G a T A	7		---------	
				N N a B N	

Non-hotspot sites

Fig. An alignment of hotspot and non-hotspot sites for the spontaneous A:T->C:G mutation spectrum in the *mutT* strain of *E.coli* (Fowler and Schaaper, 1997). The detectable position is displayed as a purine, using the appropriate DNA strand. Hotspots were inferred using the CLUSTERM program (Glazko *et al*., 1998).

Table 1. Comparison of A•T→C•G transversion in *lacI* gene from *mutT* and wild-type strains of *E. coli*

Position																				
	41	81	72	64	87	168	79	189	192	195	167	83	117	96	128	177	77	141	105	54
A•T→C•G mutations in *mutT*⁻ strain	4	10	5	2	4	9	7	23	18	10	37	5	1	5	4	7	20	5	2	8
A•T→C•G spontaneous mutations	2	3	2	2	1	2	2	8	1	6	10	0	1	0	0	3	4	0	1	3

Results of direct comparison between spectra: probability that these two spectra are different $P(\chi^2) = 0.69$ (Cariello *et al*., 1994), Kendell's ran tau correlation coefficient = 0.65 ($P < 0.01$) (Babenko, Rogozin, 1999). Positions of AA mutable motifs are underlined.

Reconstructed spontaneous mutations in human pseudogenes (Pozdniakov *et al*., 1997; *ftp.bionet.nsc.ru/pub/biology/dbms/)* were also analyzed, and the frequencies of nucleotides surrounding A•T→C•T transversions are shown in Table 2. Notably, AA and TT are the most frequent dinucleotide combinations. Such excess is statistically significant ($P(\chi^2) < 0.01$) as compared to dinucleotide frequencies in reconstructed ancestral sequences (Table 2).

A strong influence of neighboring bases was also revealed for G•C→T•A transversions, another hallmark of the 8-oxoG-dependent mutagenesis. A consensus mutable sequence GGA was derived for this type of error made *in vitro* by T4 DNA polymerase replicating 8-oxoG containing oligonucleotides (Hatahet *et al*., 1998). It was found that this nucleotide context of the 8-oxoG lesion induced less distortion of the DNA structure (Hatahet *et al*., 1998). Quite remarkably, the same GGA context for spontaneous G•C →T•A mutations in the *lacI* gene in *E. coli* was very prominent, even though DNA was replicated *in vivo* by a different replicative complex. Moreover, G•C →T•A mutations in the same context were over-represented in the collection of *p53* mutations in humans (Hatahet *et al*., 1998).

The detected context properties of mutational hotspots induced by 8-oxoG (the AA/TT motif) can be used for prediction of mutational hotspot sites in pro- and eukaryotic genes. The sequence context analysis of the lacZ gene has revealed at least one candidate mutational hotspot: T → G transversions at the position 157 result in the TGA stop-codon easily detectable by the lacZ-phenotypic selection system (Roberts, Kunkel, 1996). We have experimentally analyzed the spontaneous and induced mutations in the lacZ gene carried by the plasmid pUC19 replicated in 8-oxoG repair-deficient E.coli strains. A comparative analysis of T → G substitution frequencies at the hotspot site in a variety of DNA repair-deficient genetic backgrounds is required to assess possible contributions of 8-oxoG repair system. Preliminarily results of sequencing of several lacZ$^-$ mutants suggested that this site is a hotspot of T → G transversions.

Table 2. Frequencies of bases in position +1 and -1 in a set of spontaneous A•T → C•G transversions found in human pseudogenes (Pozdniakov *et al*., 1997)

Mutation	position -1				position +1			
	A	T	G	C	A	T	G	C
A→C					0.35	0.24	0.17	0.24
T→G	0.25	0.32	0.21	0.22				
Expected	0.25	0.22	0.22	0.31	0.26	0.23	0.29	0.22

Expected values (frequencies of AN and NT dinucleotides) were calculated in ancestral sequences used for reconstruction of spontaneous mutations (Pozdniakov *et al*., 1997). The differences between observed and expected frequencies of AA and TT dinucleotides were statistically significant ($P(\chi^2) < 0.01$).

In conclusion, these results suggested that mutagenesis due to 8-oxoG is significantly influenced by nearest neighboring bases and the context is quite evolutionarily stable. It is unlikely the revealed context properties are fingerprints of interactions between DNA and DNA repair/ replication/modification enzymes since context specificity of such interactions cannot be conserved for long evolutionary distances (Rogozin, Pavlov, 2003). Thus, the revealed context properties might reflect intrinsic properties of interactions between 8-oxoG and DNA. The AA context is the most ancient mutable motif and can be used as a fingerprint of oxidative mutagenesis in all domains of life.

Acknowledgements

We thank Nikolay Kolchanov and Yuri Pavlov for stimulating discussions and helpful comments. This work was supported by RFBR (grant No. 02-04-48342, 02-04-49889).

References

Adelman R., Saul R.L., Ames B.N. Oxidative damage to DNA: relation to species metabolic rate and life span // Proc. Natl Acad. Sci. USA. 1988. V. 85. P. 2706–2708.

Ames B.N. Endogenous DNA damage as related to cancer and aging // Mutat. Res. 1989. V. 214. P. 41–46.

Babenko V.N., Rogozin I.B. Use of a rank correlation coefficient for comparing mutational spectra // Biofizika. 1999. V. 44. P. 632–638.

Berikov V.B., Rogozin I.B. Regression trees for analysis of mutational spectra in nucleotide sequences // Bioinformatics. 1999. V. 15. P. 553–562.

Cariello N.F., Piegorsch W.W., Adams W.T., Skopek T.R. Computer program for the analysis of mutational spectra: application to p53 mutations // Carcinogenesis. 1994. V. 15. P. 2281–2285.

Drake J.W., Baltz R.H. The biochemistry of mutagenesis // Annu. Rev. Biochem. 1976. V. 45. P. 11–37.

Fowler R.G., Schaaper R.M. The role of the mutT gene of Escherichia coli in maintaining replication fidelity // FEMS Microbiol. Rev. 1997. V. 21. P. 43–54.

Glazko G.V., Milanesi L., Rogozin I.B. The subclass approach for mutational spectrum analysis: application of the SEM algorithm // J. Theor. Biol. 1998. V. 192. P. 475–487.

Hatahet Z., Zhou M, Reha-Krantz L.J., Morrical S.W., Wallace S.S. In search of a mutational hotspot // Proc. Natl Acad. Sci. USA. 1998. V. 95. P. 8556–8561.

Maki H., Sekiguchi M. MutT protein specifically hydrolyses a potent mutagenic substrate for DNA synthesis // Nature. 1992. V. 355. P. 273–275.

Michaels M.L., Miller J.H. The GO system protects organisms from the mutagenic effect of the spontaneous lesion 8-hydroxyguanine (7,8-dihydro-8-oxoguanine) // J. Bacteriol. 1992. V. 174. P. 6321–6325.

Michaels M.L., Tchou J., Grollman A.P., Miller J.H. A repair system for 8-oxo-7,8-dihydrodeoxyguanine // Biochemistry. 1992. V. 31. P. 10964–10968.

Oller A.R., Schaaper R.M. Spontaneous mutation in Escherichia coli containing the dnaE911 DNA polymerase antimutator allele // Genetics. 1994. V. 138. P. 263–270.

Pavlov Y.I., Minnick D.T., Izuta S., Kunkel T.A. DNA replication fidelity with 8-oxodeoxyguanosine triphosphate // Biochemistry. 1994. V. 33. P. 4695–4701.

Pozdniakov M.A., Rogozin I.B., Babenko V.N., Kolchanov N.A. Neighboring base effect on emergence of spontaneous mutations in human pseudogenes // Dokl. Akad. Nauk. 1997. V. 356. P. 566–568.

Roberts J.D., Kunkel T.A. Eukaryotic DNA replication fidelity // DNA Replication in Eukaryotic Cells: Concepts, Enzymes and Systems / Ed. M.D. Pamphilis. Cold Spring Harbor Laboratories, Cold Spring Harbor, New York, 1996. P. 217–247.

Rogozin I.B., Pavlov Y.I. Theoretical analysis of mutation hotspots and their DNA sequence context specificity // Mutat. Res. 2003. V. 544. P. 65–85.

Tassotto M.L., Mathews C.K. Assessing the metabolic function of the MutT 8-oxodeoxyguanosine triphosphatase in Escherichia coli by nucleotide pool analysis // J. Biol. Chem. 2002. V. 277. P. 15807–15812.

THE ARGO_SITES: AN ANALYSIS AND RECOGNITION OF THE TRANSCRIPTION FACTOR BINDING SITES BASED ON SETS OF DEGENERATE OLIGONUCLEOTIDE MOTIFS

***Vishnevsky O.V.*[1], *Ignatieva E.V.*[1], *Arrigo P.*[2]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] ISMAC, via De Marini 6 16149 Genova, Italy

* Corresponding author: e-mail: oleg@bionet.nsc.ru

Keywords: *transcription factor binding site recognition, the structure and function of the regulatory gene regions in eukaryotes, oligonucleotide motifs*

Summary

Motivation: The transcription factor binding sites (TFBS) are the critical determinants of the assembly and positioning of the basal transcription complex, thereby providing the tissue- and stage-specificity of transcription in the eukaryotic genes. Development of new methods to analyse and recognize the TFDSs are required for clarifying the structural – functional organization of the regulatory regions of the eukaryotic genes and improving our understanding of the mechanisms underlying genetic information expression.

Results: Based on the method we previously designed for identifying sets of degenerate oligonucleotide motifs, 17 TFBS samples of eukaryotes were analyzed. For the analyzed samples, sets of significant motifs were built, and the specific features of the nucleotide context at the 5'- and 3'- flanking regions were identified for the sites. The ARGO_Sites program for the recognition of the identified TFBSs on the basis of the identified motifs was developed.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/argosites/

Introduction

Assembly of the basal transcription complex, the tissue- and stage-specific transcription of the eukaryotic genes are dependent upon the presence of the transcription factor binding sites (TFBS) in the regulatory regions of the eukaryotic genes (Ignatieva *et al*., 1997). Accurate recognition of the TFBSs is required for identifying the regulatory gene regions and broadening our understanding of their structural and functional features. Most current methods for the analysis and the recognition of the TFBSs rely on multiple alignment of experimentally derived TFBSs and on building positional weight matrices or consensusi (Bucher, 1999).

The previously developed method (Vishnevsky *et al*., 2003) was applied in the analysis and recognition of the TFBSs. It is based on the identification of oligonucleotide motifs with the following properties in the examined TFBS samples: 1) degeneracy, the oligonucleotide motifs described using the expanded IUPAC alphabet; 2) region-specificity, i.e., the preferential occurrence in a particular TFBS region; 3) quasiinvariance, i.e., the occurrence in the majority of the TFBSs in a group under consideration; 4) contrast, i.e., high occurrence frequency of a particular TFBS region; 5) statistical significance determined by the binomial test. The method requires neither preliminary alignment of the sequences of interest nor additional information on them. An Internet available program for recognizing the TFBSs, ARGO_SITES (http://wwwmgs.bionet.nsc.ru/mgs/programs/argosites/) was developed from the identified oligonucleotide sets.

Results and Discussion

Search for the degenerate oligonucleotide motifs in the sequences of the SF1 binding sites. Using the proposed method, we performed a contextual analysis of a sample of the SF1 binding sites extracted from the TRRD Database (Kolchanov *et al*., 2002). The sample contained 41 nucleotide sequences of the SF1 binding sites in a region spanning from –50bp to +50bp relative to the site center. The SF1 (steroidogenic factor 1) transcription factor is a member of the nuclear receptor family, and it has a DNA binding domain of the Cys4 zinc finger type. SF1 is involved in the activation of the steroidogenic genes by interacting with DNA in a monomeric form. The question raised was, to what extent is the nucleotide context similar in the sequences containing the SF1 site in a direct and a reverse orientation? In an attempt to answer the question, we searched for the degenerate oligonucleotide motifs in a sample containing SF1 in a direct orientation and in another sample containing it in a reverse orientation.

Lists of the identified oligonucleotide motifs exemplifying the different localizations in the SF1 binding site are given in Tables 1 and 2. From a survey of the tables it is apparent that the majority of the motifs of the two samples lies in the [-10;+10] region relative to the site center. The motifs identified in this stretch partly conform to the consensus in a (GTCAAGGTCA) direct orientation or in a (TGACCTTGAC) reverse orientation. However, the flanking [-40;-10] and [+10;+40] regions relative to the site center contain different oligonucleotide motifs. Furthermore, the motifs markedly differ by the distribution of the number of motifs found in the flanking regions (Fig.).

Table 1. A characterization of oligonucleotide motifs in the sequences of the SF1 binding sites in a direct orientation

Motif	Motif location in the SF1 sample	Occurrence frequency in the SF1 site sample	Random occurrence probability of a motif
WYTNYCAS	-45: -25	0.36	10^{-9}
CNGSMNCT	-30: -10	0.36	10^{-9}
NYCAAGGY	-10: +10	0.68	10^{-31}
RAGGTCMN	-10: +10	0.31	10^{-31}
AAGGTCNN	-5: +15	0.45	10^{-18}

Table 2. A characterization of oligonucleotide motifs in the sequences of the SF1 binding sites in a reverse orientation

Motif	Motif location in the SF1 sample	Occurrence frequency in the SF1 site sample	Random occurrence probability of a motif
GGNGGAGG	-50: -30	0.21	10^{-8}
KKKGNGAG	-50: -30	0.30	10^{-8}
NRDCCTTG	-10: +10	0.69	10^{-30}
CCTTGWCN	-10: +10	0.52	10^{-22}
YCYRGRKN	+20: +40	0.47	10^{-11}
RYYCWGGN	+25;+45	0.34	10^{-9}

The sample of sites in a direct orientation contains a modest number of the degenerate motifs in the 5'- flanking regions relative to the site center. The sample of the SF1 sites in a reverse orientation contains the degenerate motifs, both in the 5'- and 3'- flanking regions. The number of identified motifs in the [+30;+40] region is comparable to the one for the site central region. An example of a degenerate motif identified in the 5'- flanking region in this sample is the GG abundant GGNGGAGG. Motifs, such as YCYRGRKN и RYYCWGGN, were identified in the 3'- flanking regions of the SF1 site.

The specificities of the oligonucleotide composition of the 3'- flanking region of the reverse SF1 sites are explainable by the presence of some other TFBS. To verify the assumption, we compared the identified motifs with the known TFBSs extracted from the TRRD Database. It proved that the majority of the identified motifs in the [+30;+40] region of the SF1 binding site share contextual similarity with the TFBS NF1.These results are in agreement with the previous (Busygina *et al.*, 2003) indicating that TF NF1 is also involved in the regulation of the steroidogenic gene expression.

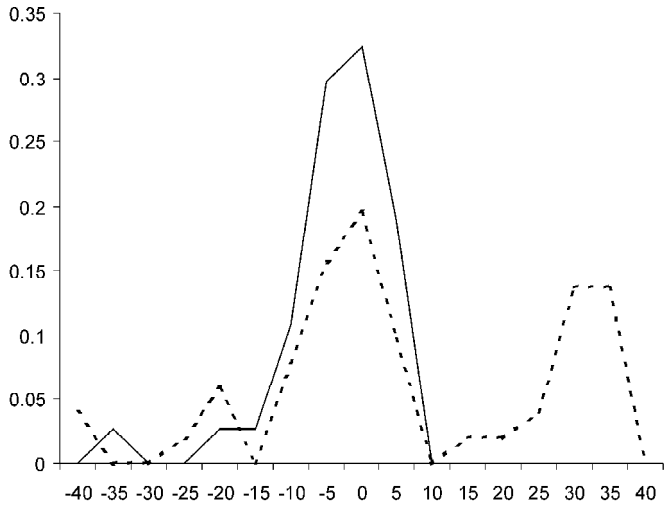


Fig. Distribution of the number of identified oligonucleotide motifs along the SF1 site sequences in a direct (solid line) and in a reverse (dotted line) orientation. The positions of the sample sequences relative to the site center are plotted along the abscissa, the proportion of motifs identified in the examined regions is plotted along the ordinate.

It appears likely that the presence of TF NF1 in the 3'- flanking region of SF1 in a reverse orientation is required for the adequate function of SF1.

Analysis and recognition of the transcription factor binding sites in eukaryotes. In the same way, 16 other TFBSs deposited in the TRRD Database were analysed. The samples also contained samples of unaligned multi-core TFBSs, such as STAT, PPRE, NFkB и COUP. For all the examined samples, we obtained sets of significant oligonucleotide motifs and developed methods for recognizing their TFBSs.

Table 3 gives the estimates for the recognition accuracy of the TFBSs at the optimum values of the recognition function parameters used to minimize type I errors. The type I errors for the different sites varied from 0.05 for the CLOCK_BMAL1 site to 0.73 for the Mono_GATA1 and Sp1 sites.

It should be noted that samples of larger size, as a rule, had a higher number of identified motifs and level of type I errors. However, the sample of the SF1 binding sites of about the same size as the E2F sample (41 and 38 sequences, respectively) had a threefold smaller number of identified motifs (66 and 187, respectively). For the NFkB sample of 47 sequences containing 142 identified motifs, type I error was twofold smaller (0.34) than the one for the HNF4 sample, consisting of 31 sequences and having 12 identified motifs and type I error estimated as 0.67.

We suggest that the marked differences in the number of identified oligonucleotides between samples of the same size may be due to the considerable heterogeneous context of the sample. This explanation may be true for the different estimates for the type I errors admitted during site recognition. The greater number of identified motifs may be due to the higher significance levels for the identified motifs.

Table 3. Estimation of the recognition accuracy of the TFBSs in eukaryotes using the proposed method based on degenerate oligonucleotide motifs

TFBS	Number of sequences in the sample	Number of identified oligonucleotide motifs	Type I error	Type II error
CLOCK_BMAL1	15	19	0.05	< 0.0002
Pu1	23	56	0.07	< 0.0002
MyoD	13	15	0.08	< 0.0002
PPRE	19	30	0.09	< 0.0002
SRF	12	22	0.12	< 0.0002
JunD1	19	29	0.13	< 0.0002
SF1	41	66	0.23	< 0.0002
ER	18	15	0.28	< 0.0002
NFkB	47	142	0.34	< 0.0002
NF1	30	52	0.36	< 0.0002
STAT	28	70	0.41	< 0.0002
HSF1	37	150	0.49	< 0.0002
COUP	32	84	0.50	< 0.0002
E2F	38	187	0.52	< 0.0002
HNF4	31	120	0.67	< 0.0002
Mono_GATA1	45	177	0.73	< 0.0002
Sp1	245	591	0.73	< 0.0002

Analysis of the data in the table showed that type I error is smaller than 0.35 for 9 of 17 samples, it is smaller than 0.55 for 5, and smaller than 0.73 for 3 samples.

This accuracy means that half of the TFBSs can lose not more than 35 % of real sites during the recognition process. In a random nucleotide sequence there are not more than 1–2 falsely recognized sites of the same TFBSs per 10,000 bp. Consequently, the method proposed for TFBS recognition is very accurate.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 01-07-90376-в, 02-07-90355, 03-07-96833-p2003, 03-07-96833, 03-07-90181-в, 02-04-48802-a, 03-04-48829, 03-07-06078-мас, 03-04-48555-a, 03-07-06082-мас); Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”; the Siberian Branch of the Russian Academy of Sciences (integration project No. 119); Russian Ministry of Industry, Science, and Technologies (grants Nos. 43.073.1.1.1501 and 43.106.11.0011, subcontract No. 38/2004); NATO (grants LST.CLG 979815).

References

Bucher P. Regulatory elements and expression profiles // Curr Opin Struct Biol. 1999. V. 9(3). P. 400–7.

Busygina T.V., Ignatieva E.V., Osadchuk A.V. Transcription regulation of genes controlling Biosynthesis of steroid hormones: description in ES-TRRD Database // Successes of Modern Biology. 2003. V. 123(4). P. 364–382. (In Russ).

Ignatieva E.V., Merkulova T.I., Vishnevskii O.V., Kel A.E. Transcription regulation of lipid metabolism genes as described in the TRRD database // Mol. Biol. (Mosc.). 1997. V. 31. P. 575–591

Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. Transcription regulatory regions database (TRRD): its status in 2002 // Nucleic Acids Res. 2002. V. 30(1). P. 312–7.

Vishnevsky O.V., Anan’ko E.A., Ignatieva E.V., Podkolodnaya O.A, Stepanenko I.V. Argo_viewer: a package for recognition and analysis of regulatory elements in eukaryotic genes // Bioinformatics of genome regulation and structure / Ed. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston; Dordrecht; London, 2003. P. 71–81.

MODELING OF CONTEXT-DEPENDENT CONFORMATIONAL PARAMETERS OF DNA DUPLEXES

Vorobjev Y.N., Emelianov D.Y.

Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, 630090, Lavreniev Ave. 8; e-mail: ynvorob@niboch.nsc.ru

Keywords: *molecular dynamics, DNA conformational dynamics, context dependent DNA conformational parameters*

Resume

Motivation: A context dependent deformability of DNA can be itself a significant factor of a binding site recognition by a DNA binding protein. An extraction of a context-dependent deformability parameters from static X-ray structures of crystals of DNA duplexes can includes artifacts due to crystal packing. A molecular dynamic simulation of a series of 14-base pairs DNA duplexes in water solution with counter ions have been done. The context-dependent conformational parameters are extracted from the simulated trajectories of atomic thermal fluctuations of DNA atoms.

Results: molecular modeling provide data to expand bioinformatics data bases beyond the capacity of experimental methods and provide a new knowledge.

Introduction

Formation of the protein-DNA complex is a complicated process which depends on the three dimension structure and conformational dynamics of DNA binding site. A context dependent deformability of DNA can be itself a significant factor of a binding site recognition by a DNA binding protein. Therefore a reliable set of context dependent deformability parameters of DNA can serve as a natural descriptors to identify DNA protein binding sites. Unfortunately a complete set of context dependent dynamic deformability parameters is still unknown because analysis of static crystal structures of DNA duplexes or protein/DNA complexes can include artifacts due to crystal packing effect which can be significant for the DNA duplexes [1]. A recent progress in the macromolecular modeling methods gives a tools to obtain a simulated bioinformatics data in aqueous solution to extract a knowledge we need [1]. Molecular modeling of internal conformational dynamics of DNA duplexes due to thermal fluctuation is able to provide data of the context dependent deformability. The molecular dynamics simulations of eight 14-base pairs DNA duplexes of 1.5 nano sec length for each are performed in an aqueous solvent with neutralized counterions. A dynamic average and value of thermal fluctuations of conformational parameters which defines dynamic deformability for all possible ten types of pair nucleotide steps are extracted via statistical analysis of molecular dynamics data.

Methods

A double stranded DNA duplex of A,G,C,T nucleotides can be represented as a linear sequence of ten types of pair-nucleotide steps [1]. A sequences of eight DNA duplexes of 14-base pairs are selected for simulations. This sequences includes all ten types of dinucleotide steps with a different flanking sequences:

dna14-1 : (ACGTTGAACGACTG); dna14-2 : (GTCAGTAAGTGCAG);
dna14-3 : (TGTCGGAATGCTAC); dna14-4 : (GTCATCGGCTGCTA);
dna14-5 : (GGTTAATATGCCGA); dna14-6 : (AGTATATAAAACGC);
dna14-7 : (GGCTTAGGTAATTG); dna14-8 : (TTAGGCTTCGGCCA).

Molecular dynamics simulations have been done with amber6 [2] program using param98 force-field parameter set. Simulation protocol consist of the next stages: 1) calculation of initial coordinates of duplex atoms in standard B-form; 2) short energy minimization in vacuum with distant dependent dielectric constant; 3) solvatation of the dna-duplex in the rectangular box with 9 E distance from the nearest dna atom to the box side; 4) neutralization of the dna-duplex by 26 Na^+ ions; 5) energy minimization of water and ion positions until grad(E) < 0.1 kcal/mol/ Å; 6) slow heating from 1 to 300 K in 10 ps with the soft harmonic restraint potential (the harmonic potential constant, $K_h = 0.1$ kcal/mol/$Å^2$) for dna-duplex atoms; 7) final equilibration during 50 ps with the soft harmonic restraint potential for atoms of dna-duplex flanking base pairs; 8) productive molecular dynamics run of 1500 ps at T = 300K, P = 1bar with PME for the long-range electrostatic forces and weak harmonic restraint potential (the harmonic potential constant, $K_h = 0.02$ kcal/mol/ $Å^2$) and trajectory collection with 1 ps interval.

Several programs have been developed for processing of the molecular dynamical trajectories generated by the amber6 program. Program extractPDBfromAmberTra extracts a series of atomic coordinate files of the dna duplex in the pdb format from the amber6 trajectory.

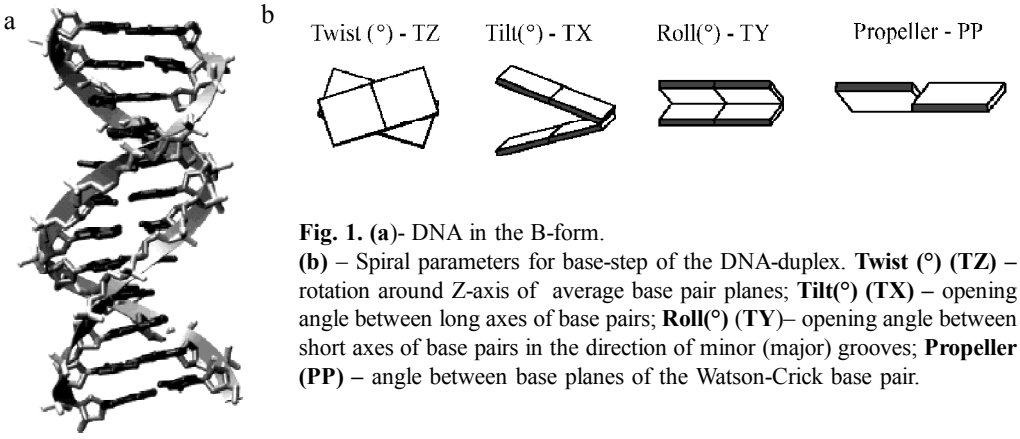


Fig. 1. (a)- DNA in the B-form.
(b) – Spiral parameters for base-step of the DNA-duplex. **Twist (°) (TZ) –** rotation around Z-axis of average base pair planes; **Tilt(°) (TX) –** opening angle between long axes of base pairs; **Roll(°) (TY)**– opening angle between short axes of base pairs in the direction of minor (major) grooves; **Propeller (PP) –** angle between base planes of the Watson-Crick base pair.

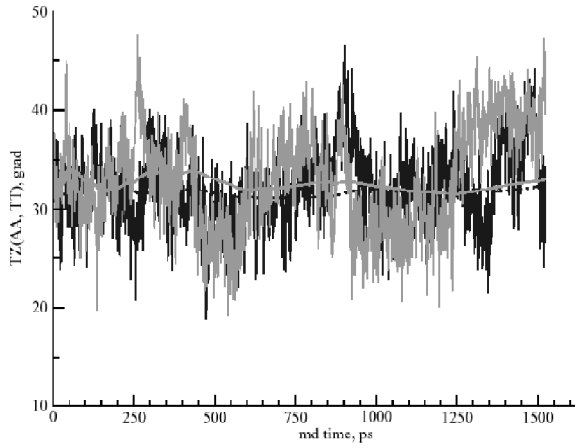


Fig. 2. Fluctuation of the spiral angle Twist (TZ) for the T4T5 (blue) and A7A8 (green) nucleotide step of the DNA 14-1 duplex.

Program spiralParDna calculates the spiral parameters [3] for each base-pair step of the DNA duplex, see Fig. 1.
A typical fluctuation behavior of the base step spiral parameters along a molecular dynamic trajectory of the dna14-1 duplex are shown in Fig. 2. It can be seen that high frequency fluctuations (several ps scale) have a large amplitude and describe a fast local conformational fluctuations. The average, over the 50 ps window, shows quite smooth behavior with a period of slow fluctuation of about 500–600 ps. Therefore it can be concluded that trajectory of 1500 ps of length provides a reasonable amount of data to obtain an average and statistical fluctuation values for the spiral parameters of DNA duplex in water solution. A similar duration of molecular dynamic trajectories have been used in ref. 4.

Table. Simulated spiral parameters and its standard deviations for ten dinucleotide steps

n - type dinucleotide step 5'-3'	K [f]	<TX>	<TY>	<TZ>	<PP>[a]	$<\Delta TX^2>^{1/2}$	$<\Delta TY^2>^{1/2}$	$<\Delta TZ^2>^{1/2}$	$<\Delta PP^2>^{1/2}$ [b]
1 **A-A**	13	3.5	-6.2	33.0	-20.1 [a]	6.4	10.3	4.7	15.8 [b]
		1.0	***4.9***	***2.2***	***4.3*** [c]				
2 **A-T**	9	-0.2	-1.5	32.0	-19.2 [a]	7.0	9.8	4.0	15.9 [b]
		2.1	***4.8***	***3.5***	***5.4*** [c]				
3 **A-G**	10	5.6	-3.1	35.0	-14.6 [a]	6.4	10.4	4.2	17.0 [b]
		1.6	***5.0***	***2.9***	***8.5*** [c]				
4 **A-C**	10	2.6	-1.2	33.2	-17.1 [a]	6.9	9.5	4.1	16.2 [b]
		1.5	***6.0***	***2.4***	***7.5***				
5 **T-A**	8	0.1	-15.1	32.5	-18.4 [a]	6.4	10.6	5.5	17.0 [b]
		1.8	***6.3***	***2.9***	***5.3*** [c]				
6 **T-G**	8	1.7	-15.4	30.2	-12.3 [a]	5.6	10.5	5.3	16.9 [b]
		1.6	***4.9***	***3.8***	***5.8*** [c]				
7 **T-C**	7	-1.5	-0.5	35.7	-14.2 [a]	6.3	10.6	4.9	16.5 [b]
		1.2	***3.5***	***2.5***	***10.0*** [c]				
8 **G-G**	7	1.9	-1.8	33.8	-13.2 [a]	6.1	10.3	4.9	16.6 [b]
		1.7	***1.1***	***2.1***	***4.3*** [c]				
9 **G-C**	8	1.0	3.5	36.5	-13.9 [a]	6.9	9.2	4.0	16.1 [b]
		0.9	***2.3***	***1.2***	***2.9*** [c]				
10 **C-G**	7	-0.3	-14.9	31.9	-8.4 [a]	6.0	10.3	5.8	18.4 [b]
		2.9	***6.5***	***7.1***	***6.3*** [c]				
Average over all 10 steps [e]	10	1.4	-5.6	33.4	-15.1	6.4	10.2	4.7	16.6 [d]
		2.0	***6.6***	***1.8***	***3.4***	***0.4***	***0.5***	***0.6***	***0.7*** [g]

[a] – average values over different flanking DNA sequences; [b] – values of thermal fluctuations of spiral parameters averaged over different flanking DNA sequences; [c] – standard deviations for average spiral parameters over different flanking DNA sequences; [d] – average values over all types of dinucleotide steps; [g] – standard deviations for over 10 types of diducleotide steps; [e] – average values over all dinucleotide step for all flanking sequences; [f] – number of different flanking sequences.

Results

Table shows results of simulation of spiral parameters for the ten types of dinucleotide steps. It can be seen that the di-nucleotide step spiral parameters depends on context, i.e. of letters of the step, and they demonstrate a different dependence on flanking sequences. The most context sensitive is the average roll angle TZ (see. Fig. 1) with standard deviation of 6.6° over all contexts. The roll parameter is responsible for DNA bending into major/minor groove directions. The roll parameter has the largest value of thermal fluctuations and the one is the most soft and flexible type of conformational perturbation of the DNA duplexes. Thereby a global DNA bending in the major/minor grooves direction are context dependent and the major type of DNA deformation. A similar conclusion have been done from the crystal DNA duplex structure analysis [5–7]. The seven (numbers 1–6, 10) of the ten dinucleotide steps demonstrate a large standard deviation of average spiral parameters over different context of flanking sequences, 4.8 – 7.1°. I.e. a minimal context which more precisely defines the local spiral parameters of dinucleotide step **A-B** is not the step context itself but a quartet X-**A-B**-Y of nucleotides. Quartet based analysis needs more simulations and statistical which are in progress.

Acknowledgements

The work was supported in part by Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501).

References

1. Vorobjev Y.N. In silico modeling and conformational mobility of DNA duplexes // Mol. Biol. 2003. V. 37, N 2. P. 210–222. (Transl. from Rus.).
2. URL: 2001. AMBER6 Home page: http://www.amber.ucsf.edu/amber/index.html.
3. Olson W.K., Lu X.-J., Westbrook J., Berman H.M., Bansal M., Burley S.K., Dickerson R.E., Harvey S.C., Heinemann U., Neidle S., Wolberger S.K. A standard reference frame for the description of nucleic acid base-pair geometry // J. Mol. Biol. 2001. V. 313. P. 229–237.
4. Qian X., Strahs D., Schlic T. Dynamic simulations of 13 TATA variants refine kinetic hypothesis of sequence/activity relationships // J. Mol. Biol. 2001. V. 308. P. 681–703.
5. Olson W.K., Gorin A.A., Lu X.-J., Hock L.M., Zhurkin V.B. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes // Proc. Natl Acad. Sci. USA. 1998. V. 95. P. 11163–11168.
6. Goodsell D.S., Kaczor-Grzeskowiak M., Dickerson R.E. The crystal structure of C-C-A-T-T-A-A-T-G-G: implications for bending of B-DNA at T-A steps // J. Mol. Biol. 1994. V. 239. P. 79–96.
7. Berman H.M. Crystal Studies of B-DNA: TheAnswers and the Questions Biopolymers // 1997. V. 44. P. 23–44.

USE OF AN INTEGRATED RULE SYSTEM FOR IDENTIFICATION OF THE TRANSCRIPTION FACTOR BINDING SITES FOR MCM1 AND FKH2 IN *S. CEREVISIAE*

***Walker N.J.*[1,2], *Sharrocks A.D.*[1], *Attwood T.K.*[1,2]**

[1] School of Biological Sciences, University of Manchester U.K.; [2] Department of Computer Science. University of Manchester U.K.
* Corresponding author: e-mail: walker@bioinf.man.ac.uk

Keywords: *Mcm1, Fkh2, Saccharomyces cerevisiae, transcription factor, binding site, composite elements, comparative genomics*

Summary

Motivation: To better understand the co-ordinate activity of two transcription factors, Mcm1 and Fkh2, we wish to study the sequence variation in their composite binding sites. This, we hope, will improve the in silico location of their sites on a genome wide scale. In particular, our approach is to avoid the inherent bias of sparse weight matrices towards strong (less variable) binding sites. We follow a strategy that seeks to take the emphasis off the sequence itself, and instead, produce a set of putative binding sites through the application of rules based on distances between composite motifs, conservation within closely related species, and gene annotation.

Results: A system has been developed that integrates yeast species alignment data, annotation data and genome wide location analysis experiment data. This allows formulation of several rules, giving a set of potential binding sites that include all generally accepted composite Mcm1 and Fkh2 binding sites. With this approach several further potential binding sites are uncovered.

Introduction

The transcription factors Mcm1 and Fkh2 have been implicated in the co-ordinated regulation of genes involved in G2-M phase of the cell cycle. It is interesting to use bioinformatics approaches to further assess the binding location of these transcription factors and test the boundaries of this hypothesis. A recent review by Wasserman and Sandelin (2004) has suggested reasons why a computational search model, based only on functional transcription factor binding sites, leads to many false positives. They note that there are often limited functional binding site sequences available to parameterise the model, and that such sites may include a significant amount of variation, making them difficult to identify. Such is the case with Mcm1 and Fkh2. Core sequences from known Mcm1 binding sites are variable (Shore, Sharrocks, 1995), while partner Fkh2 binding sites are less variable but are shorter.

Methods and Algorithms

Our approach is to try to capture weaker sites, which are not well represented in the known set of transcription factor binding sites. To achieve this, the following steps are employed:

1. Consensus sequences:

Consensus sequences with mismatches are known to be problematic for obtaining good specificity (Stormo, 2000); however we use them here to establish a large amount of variation in the represented binding sites. Composite sites, are retrieved that match the wide consensus pattern with one mismatch.

2. Application of Rules:

a) *Spacing between composite motifs*: based on information in Boros *et al*. (2003). This parameter is set to vary between 3 and 20 base pairs.

b) *Conservation between closely related species*: the average conservation within the site, over the average conservation over 50 base pairs surrounding the site should be greater than 0.9.

c) *Annotation markers*: results can then be enhanced or further specified by indicating key words from annotation associated with a gene. In particular, marking genes that are known to be cell cycle regulated is useful for these particular transcription factors.

3. Comparison with Genome Wide Location Analysis (GWLA) Experiment:

The resulting set of potential sites is analysed against joint p-value scores for Mcm1 and Fkh2 transcription factor binding sites produced by a GWLA experiment (Lee *et al*., 2002).

Implementation and Results

Execution of the method requires a program to produce all matching instances and their positions, and a relational database combining, this data with promoter regions over *S. cerevisiae, S. bayanus, S. paradoxus,* and *S. mikitae* (Kellis *et al*., 2003) and annotation data gained www.yeastgenome.org, GWLA data.

Based on these rules, the set of matches is greatly reduced (Table 1), making it possible to study the effect of different parameters on known and unknown sites.

Comparison of the joint probability of Mcm1 and Fkh2 binding together on a promoter given by GWLA reveals that 30 of these 38 sites occur in promoters identified to have a joint p-value of less than 0.05 (see Table 2).

Table 1. Rule based reduction of potential binding sites

Step	No binding site matches
1.	MCM1 ~ 30000 FKH2 ~ 50000
2. Rule a	~ 8000
2. Rule b	260
2. Rule c	38

All sites that were strongly identified by Simon *et al.* (2001) are included in these results. Without filtering on genes involved in the cell cycle, at a p-value of 0.0062, (the highest p-value of a known independently verified gene in Table 2), the GWLA of Lee *et al*. (2002) returns 51 sites. Without any rules applied to potential sequences in the promoter, 441 promoter regions are returned at this p-value, clearly involving many false positives. At 0.02, the lowest p-value in Table 2, the figures are respectively, 136 and 689 sites. This shows that Mcm1 and Fkh2 do bind cooperatively to weak sites, and underscores the problem of identifying such weak sites on a genome wide scale.

The system can be easily extended to employ further rules, such as conditional rules on nucleotides within the sequence, and rules based on alignment gaps within a sequence. However, after application of rules a to c, they lead only to trivial reduction in the resulting data set.

Discussion

We show that by emphasising non-sequence characteristics of binding sites using simple rules, it is possible both to capture strong sites, and to suggest putative weaker binding sites. Clearly, this method relies on well chosen initial parameters to achieve reductions in matches with out loss of known functional binding sites. This underscores the fact that detection of weak binding sites is problematic for both computational and wet-lab techniques when applied at a genome-wide scale. We are currently using this rule based system to study variation in binding sites given such rules, ultimately with a view to automating parameter setting. Our analyses will also produce test-sets for discrimination algorithms. Several limiting factors in this approach include the reliance on good multiple alignments of related genomes, and the assumption that most binding sites, whether weak or strong, are conserved in closely related species.

Table 2. 30 putative composite Mcm1, Fkh2 binding sites identified using rules a to c. Underlined ORF codes indicate that the gene or site has previously been implicated in binding by the two transcription factors

Gene name	Alias	ORF code	MCM1 motif	FKH2 motif	Joint p-value for MCM1 and FKH2 binding to promoter given by GWLA study
CDC20	PAC5	YGL116W	gccgaaagagg	gtaaata	2.4E-18
SPO12		YHR152W	tcctaatttgg	gtaaaca	1.56E-16
BUD4		YJR092W	acccgatttgg	gtaaaca	1.078E-15
SWI5		YDR146C	acctgtttagg	gtaaaca	2.016E-15
		YJL051W	tccttttggg	gtaaaca	5.27E-15
		YJL051W	ttccttttgg	gtaaaca	5.27E-15
		YLR190W	tcccaaacggg	ataaata	1.558E-13
		YLR190W	acctattttat	gtaagca	1.558E-13
CLB2		YPR119W	accgaatcagg	gtcaaca	8.84E-13
SUR7		YML052W	accccaatcgg	gtaaaca	9.96E-13
SUR7		YML052W	ccccaatcggg	gtaaaca	9.96E-13
UTH1		YKR042W	tcctgattcgg	gtgaata	9.36E-12
KIN3	FUN52|NPK1	YAR018C	ttcctaatta-g	gtaaata	8.4E-10
ALK1		YGL021W	gccctttttgg	gtaaaca	1.218E-08
DBF2		YGR092W	ccctttttagg	gtaatta	6.12E-07
DBF2		YGR092W	gccctttttag	gtaatta	6.12E-07
NCE102	NCE2	YPR149W	t-cctttataaa	gtaatta	0.00000273
		YNL058C	tcctaaagcgg	gtaaaca	4.361E-06
		YNL058C	acccttaatat	ggaaata	4.361E-06
PMA1		YGL008C	acctttttccgg	gaaaaca	0.00000492
PMA1		YGL008C	gcctcaataag	gtaa--atc	0.00000492
		YMR253C	tcctaactagg	gtaaaaa	0.0000976
		YMR253C	ttcctaactag	gtaaaaa	0.0000976
BNS1		YGR230W	tcctaaaaagg	gtaatta	0.000414
BNS1		YGR230W	atcctaaaaag	gtaatta	0.000414
MYO1		YHR023W	gcccaaaaggg	gtaaaca	0.0013
SUT1		YGL162W	t--cctaatacgg	gtaaacc	0.00224
HOF1	CYK2	YMR032W	tcctctttggg	gtaaaca	0.0062
		YPR013C	gccatttacgg	ggaaata	0.02553

Acknowledgements

We are grateful to the Wellcome Trust for individual support(NW).

References

Boros J., Lim F.L., Darieva Z., Pic-Taylor A., Harman R., Morgan B.A., Sharrocks A.D. Molecular determinants of the cell-cycle regulated Mcm1p-Fkh2p transcription factor complex // Nucleic Acids Res. 2003. V. 31(9). P. 2279–2288.

Kellis M., Patterson N., Endrizzi M., Birren B., Lander E. S. Sequencing and comparison of east species to identify genes and regulatory elements // Nature. 2003. V. 423(6937). P. 241–254.

Lee T.I., Rinaldi N.J., Robert F., Odom D.T., Bar-Joseph Z., Gerber G.K., Hannett N.M., Harbison C.T., Thompson C.M., Simon I., Zeitlinger J., Jennings E.G., Murray H.L., Gordon D.B., Ren B., Wyrick J.J., Tagne J.B., Volkert T.L., Fraenkel E., Gifford D.K., Young R.A. Transcriptional regulatory networks in *Saccharomyces cerevisiae* // Science. 2002. V. 298(5594). P. 763–764.

Shore P., Sharrocks A.D. The MADS-box family of transcription factors // Eur. J. Biochem. 1995. V. 229(1). P. 1–13.

Simon I., Barnett J., Hannett N., Harbison C.T., Rinaldi N.J., Volkert T.L., Wyrick J.J., Zeitlinger J., Gifford D.K., Jaakkola T.S., Young R.A. Serial regulation of transcriptional regulators in the yeast cell cycle // Cell. 2001. V. 106(6). P. 697–708.

Stormo G. DNA binding sites: representation and discovery // Bioinformatics. 2000. V. 1. P. 16–23.

Wasserman W.W., Sandelin A. Applied bioinformatics for the identification of regulatory elements // Nature Reviews Genetics. 2004. V. 5. P. 276–287.

A PECULIAR CODON USAGE PATTERN REVEALED AFTER REMOVING THE EFFECT OF DNA METHYLATION

Xuhua Xia

Department of Biology, University of Ottawa; e-mail: xxia@uottawa.ca

Keywords: *DNA methylation, deamination, codon usage, genome, genomics*

Summary

DNA methylation and deamination increases the C→T mutation rate in CpG dinucleotides, especially in vertebrate genomes. This has profound effect on codon usage in heavily vertebrate genomes, and may obscure the effect of other factors on codon usage bias. We have classified the sense codons into three groups: those decreased by DNA methylation (i.e., CpG-containing codons), those increased by DNA methylation (i.e., TpG- and CpA-containing codons), and those not directly affected by DNA methylation, and studied the codon usage of the last group. RRR and YYY codons are used significantly more frequently than the rest of the codons. This pattern is much stronger in vertebrate genomes than in other genomes and can be used as a content sensor in gene finding.

Introduction

DNA methylation is a ubiquitous biochemical process observed in both prokaryotes and eukaryotes. In vertebrates, DNA methylation is catalyzed by methyltransferases of which a typical representative is the mammalian DNMT1. DNMT1 has five domains of which the NlsD, ZnD and CatD domains bind specifically to unmethylated CpG, methylated CpG and hemimethylated CpG sites, respectively (Fatemi *et al*., 2001). Methylation of C in the CpG dinucleotide greatly elevates the mutation rate of C to T through spontaneous deamination of the resultant m^5C (Xia, 2003).

DNA methylation should have direct consequences on the evolution of codon usage. The frequency of the CpG-containing codons (e.g., CGA) should be reduced by the joint effect of DNA methylation and spontaneous deamination, and the frequency of TpG-containing and CpA-containing codons, should increase consequently. This has profound effect on codon usage. For example, four codon families (i.e., Ser, Pro, Thr, Ala) have codons with the CpG dinucleotide occupying codon positions 2 and 3. DNA methylation and spontaneous deamination will mediate the change of these NCG codons to NTG and NCA codons, with the former change being nonsynonymous and the latter synonymous. Because nonsynonymous substitution is generally much rarer than the synonymous substitutions in a large number of protein-coding genes in many organisms studied (Xia, 1998b; Xia, Hafner *et al*., 1996; Xia, Li, 1998), we should expect NCG to change to NCA more often than to NTG codons. This implies that NCG codons will be underused and NCA codons will be overused in these four codon families.

The effect of DNA methylation may obscure the effect of other factors contributing to codon usage bias. Several factors have already been postulated to affect codon usage. For example, tRNA molecules carrying the same amino acids often differ much in concentration and an increase of the synonymous codons matching the anticodon of the most abundant tRNA would increase the translation rate (Bulmer, 1991; Ikemura, 1992; Xia, 1998a). Although these two hypotheses can account for much variation in codon usage patterns, there is still much unexplained variation, much of which might be caused by the confounding effect mediated by DNA methylation and spontaneous deamination.

One method to remove the effect of DNA methylation on codon usage is simply to classify the sense codons into three groups, one including codons whose frequencies would be reduced by DNA methylation (i.e., all CpG-containing codons), one including the codons whose frequencies

would be increased by DNA methylation (i.e., all TpG- and CpA-containing codons) and the third including all other codons whose frequencies are not directly affected by DNA methylation. The codon usage pattern in this third group should not be confounded by the effect of DNA methylation on codon usage and may consequently reveal codon usage patterns not observed previously.

Materials and Methods

We have used the following vertebrate representatives: *Homo sapiens*, *Xenopus laevis*, and *Danio rerio*, and the following non-vertebrate representatives *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*. The coding sequences were extracted from UniGene files (xx.seq.uniq, where "xx" stands for the species abbreviation) available at ftp.ncbi.nih.gov/repository/UniGene for the following species: *Homo sapiens*, *Xenopus laevis*, and *Danio rerio*. The coding sequences for *D. melanogaster* were retrieved from scaffold data (AE00xxxx.ffn) available at the FTP site (ftp.ncbi.nih.gov/genbank/genomes). Only results from the first scaffold (AE002566) with 1191 CDS sequences was presented because the codon usage patterns from the other 18 scaffolds are all very similar. The *C. elegans* data were retrieved from the same FTP site. Only results from the first *C. elegans* chromosome was presented because the codon usage patterns from the other chromosomes are similar. The coding sequences for *Saccharomyces cerevisiae* were retrieved from Saccharomyces Genome Database at http://genome-www.stanford.edu/Saccharomyces. Only complete CDSs starting with a methionine codon and ending with a termination codon were used.

Codon frequencies depend on nucleotide frequencies, e.g., GC-rich DNA should have more GC-rich codons. Furthermore, the nucleotide frequencies typically are different at different codon position so that even a CDS sequence with equal number of A, C, G, and T may have codon frequencies differing much from each other. For example, if G occurs only at the first codon position and C only at the second codon position, then we would have many GCN codons but few CGN codons. Therefore, we have compiled the observed codon frequencies and computed the expected frequencies from nucleotide frequencies for each codon positions. Designating observed codon frequencies as O and expected as E, we computed the (O-E)/E which is termed SR (for standard residue) hereafter. We expect the codons in the FROM group to have small SR values and those in the TO group to have large SR values.

The counting of codon frequencies was done in two ways, one including the initiating codon and the other excluding the initiating codon. The conclusions presented in this paper are valid for both counting procedures, and only the result from the counting procedure excluding the initiating codon is presented in tables. The expected codon frequencies were calculated as follows. Let P_{ij} be the site-specific nucleotide frequencies of the sense codons, where i = 1, 2, 3 corresponding to codon positions 1, 2, and 3, and j = 1, 2, 3, 4 corresponding to A, C, G and T, respectively. The expected frequency of codon AGA is then

$$P_{AGA} = \frac{P_{1,1}P_{2,3}P_{3,1}}{SumP}, \qquad (1)$$

where SumP is the sum of the $P_{1i}P_{2j}P_{3k}$ terms for all sense codons. For the universal genetic code, SumP is the sum of 61 $P_{1i}P_{2j}P_{3k}$ terms. The calculation of P_{ij} did not include stop codons because the inclusion would underestimate the frequency of C. The computation of site-specific nucleotide frequencies, the expected codon and di-codon frequencies, and all subsequent statistical analyses were done with DAMBE (Xia, 2001; Xia, Xie, 2001).

Table 1. The observed (O) and the expected (E) frequencies for the 61 sense codons classified into three groups, with IDs of "-1", "1", and "0" for codons decreased, increased and not affected by DNA methylation, respectively. SR = (O-E)/E. Based on 15109 human CDS sequences. The "1" group does not include the UG-containing UGA codon because it is not a sense codon

Codon	O	E	SR	ID	Codon	O	E	SR	ID
ACG	44084	133005.8	-0.669	-1	CCC	144034	128711.1	0.119	0
CCG	51567	122948.6	-0.581	-1	AGA	87107	73850.7	0.180	0
CGA	45717	68266.5	-0.330	-1	GGC	163689	134313.1	0.219	0
CGC	76261	104560.8	-0.271	-1	GCC	202803	165335.3	0.227	0
CGG	83435	99879.5	-0.165	-1	AGC	140972	113114.0	0.246	0
CGU	33777	79077.3	-0.573	-1	AAG	233830	177648.4	0.316	0
GCG	54254	157933.0	-0.657	-1	CCU	128978	97341.7	0.325	0
UCG	32683	84626.2	-0.614	-1	GAG	291059	210942.2	0.380	0
GUA	52660	120106.6	-0.562	0	GGA	122104	87691.4	0.392	0
AUA	54170	101149.7	-0.465	0	UCC	125479	88592.6	0.416	0
CUA	51936	93501.3	-0.445	0	UUC	144076	98573.5	0.462	0
GUC	103360	183962.1	-0.438	0	AAA	183074	121420.6	0.508	0
GUU	80933	139126.9	-0.418	0	GAA	220405	144176.6	0.529	0
AAC	139177	185974.7	-0.252	0	UCU	109113	67000.8	0.629	0
AGG	83842	108049.7	-0.224	0	UUU	125983	74549.2	0.690	0
GGU	79894	101578.4	-0.214	0	ACA	109339	90908.0	0.203	1
GAC	185126	220829.0	-0.162	0	AUG	143532	147990.4	-0.030	1
UUA	56384	64357.4	-0.124	0	CAA	91270	112239.4	-0.187	1
CUU	95737	108308.3	-0.116	0	CAC	108309	171912.2	-0.370	1
AAU	125990	140649.0	-0.104	0	CAG	250071	164215.5	0.523	1
ACU	96179	105304.3	-0.087	0	CAU	79270	130013.8	-0.390	1
GGG	118196	128299.8	-0.079	0	CCA	125584	84033.9	0.494	1
UAC	109345	118328.1	-0.076	0	CUG	283383	136800.1	1.072	1
CUC	137174	143211.9	-0.042	0	GCA	117312	107945.4	0.087	1
AUC	148829	154926.7	-0.039	0	GUG	203095	175725.9	0.156	1
ACC	135017	139239.8	-0.030	0	UCA	88625	57841.0	0.532	1
GAU	164887	167008.6	-0.013	0	UGC	89379	71969.8	0.242	1
UAU	88939	89489.2	-0.006	0	UGG	90399	68747.6	0.315	1
AUU	117197	117168.0	0.000	0	UGU	76049	54429.3	0.397	1
AGU	90301	85545.9	0.056	0	UUG	93206	94160.3	-0.010	1
GCU	135067	125039.8	0.080	0					

Results and Discussion

The nucleotide frequencies for each of the three codon positions were counted for the 15109 human CDS sequences. These nucleotide frequencies were then used to compute the expected frequencies of the 61 sense codons for the human CDS sequences. The observed frequencies of the 61 sense codons (Table 1) are largely consistent with the effect of methylation on codon usage. The eight CpG-containing codons all have observed frequencies much lower than their respective expected frequencies. The SR value (Table 1) measures the deviation of the observed value from the expected value and is independent of sample size. The mean SR value for the eight CpG-containing codons is -0.4823. In contrast, most of the 15 UpG-containing and CpA-containing codons have their observed frequencies higher than their respective expected frequencies, with the mean SR value equal to 0.2022.

The mean SR value for those codons not affected by DNA methylation is intermediate, being 0.0494. Significant differences exist among the three means (ANOVA and multiple comparisons with $P < 0.0001$), suggesting a significant effect of DNA methylation on the usage of sense codons. Note that this effect of methylation on codon usage would be much obscured if we looked at only the observed frequencies without comparing them to the expected codon frequencies (Table 1) based on the nucleotide frequencies at the three codon positions.

The 38 codons not directly affected by DNA methylation were sorted by the SR value, which reveals a novel pattern of codon usage (Table 1). The RRR and YYY codons are used more frequently than the rest of the codons, and this pattern would have been obscured if we did not eliminate the methylation effect because the UpG-containing codons are non-RRR and non-YYY but generally have large SR values.

Designating the 16 RRR and YYY codons as Group 1 and the other 22 codons as Group 2, we found highly significant difference in the mean SR value between the two groups (Table 2). The difference is true not only for the vertebrate species in Table 2, but also for a number of other vertebrate species we tested, including *Mus musculus*, *Rattus norvegicus*, *Bos Taurus*, *Gallus gallus*, and *Alligator mississippiensis*. The difference is smaller for non-vertebrate species, and is not significant for *D. melanogaster*. However, the direction of the difference, even for *D. melanogaster*, is consistent.

Table 2. T-tests of whether RRR and YYY codons are used more frequently. Mean1 – Mean SR for all RRR and YYY sense codons, Mean2 – Mean SR for codons other than RRR and YYY. The degree of freedom is 36 for all tests

Species	N_{CDS}	Mean1 (SE)	Mean2 (SE)	T	P
H. sapiens	15109	0.2802 (0.0695)	-0.1184 (0.0495)	4.8068	<0.0001
X. laevis	1555	0.2698 (0.0642)	-0.1306 (0.0383)	5.6645	<0.0001
D. rerio	900	0.2285 (0.0952)	-0.1380 (0.0491)	3.6896	0.0007
C. elegans	2474	0.1595 (0.1313)	-0.1929 (0.0486)	2.8143	0.0079
D. melanogaster	1191	-0.0097 (0.1204)	-0.0336 (0.0660)	0.186	0.8535
S. cerevisiae	6357	0.1308 (0.0681)	-0.0565 (0.0517)	2.2325	0.0319

Our results suggest that that DNA methylation can contribute substantially to codon usage bias. Taking DNA methylation in account should greatly increase the explanatory power of the two existing hypotheses, i.e., the transcriptional (Xia, 1996) and the translational (Bulmer, 1991; Ikemura, 1992; Xia, 1998a) hypotheses on codon usage bias.

Acknowledgements

This study is supported by grants from University of Ottawa and NSERC of Canada.

References

Bulmer M. The selection-mutation-drift theory of synonymous codon usage // Genetics. 1991. V. 129. P. 897–907.

Fatemi M., Hermann A., Pradhan S. *et al*. The activity of the murine DNA methyltransferase Dnmt1 is controlled by interaction of the catalytic domain with the N-terminal part of the enzyme leading to an allosteric activation of the enzyme after binding to methylated DNA // J. Mol. Biol. 2001. V. 309. P. 1189–99.

Ikemura T. Correlation between codon usage and tRNA content in microorganisms, in Transfer RNA in protein synthesis / Eds. D.L. Hatfield, B. LeeJ. Pirtle, CRC Press: Boca Raton, Fla. 1992. P. 87–111.

Xia X. Maximizing transcription efficiency causes codon usage bias // Genetics. 1996. V. 144. P. 1309–1320.

Xia X. How optimized is the translational machinery in *E. coli*, *S. typhimurium*, and *S. cerevisiae*? // Genetics. 1998a. V. 149. P. 37–44.

Xia X. The rate heterogeneity of nonsynonymous substitutions in mammalian mitochondrial genes // Mol. Biol. Evol. 1998b. V. 15. P. 336–344.

Xia X. Data analysis in molecular biology and evolution. Boston: Kluwer Academic Publishers. 2001.

Xia X. DNA methylation and Mycoplasma genomes // J. Mol. Evol. 2003. V. 57. S21–S28.

Xia X., Hafner M.S., Sudman P.D. On transition bias in mitochondrial genes of pocket gophers // J. Mol. Evol. 1996. V. 43. P. 32–40.

Xia X., Li W.-H. What amino acid properties affect protein evolution? // J. Mol. Evol. 1998. V. 47. P. 557–564.

Xia X., Xie Z. DAMBE: Software package for data analysis in molecular biology and evolution // J. Hered. 2001. V. 92. P. 371–373.

DEVELOPMENT OF METHOD FOR *IN SILICO* MAPPING OF QUANTITATIVE TRAIT LOCI

Zykovich A.S., Axenovich T.I.*

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; Novosibirsk State University, Novosibirsk, Russia
* Corresponding author: e-mail: aks@bionet.nsc.ru

Keywords: *in silico mapping, SNP, mouse genomics, quantitative trait loci*

Summary

Motivation: A new *in silico* method for mapping of genes controlling complex traits in mice has been recently suggested. It does not require generation and analysis of experimental intercross progeny. *In silico* mapping is based on computer analysis of databases containing information about specific phenotypes and single nucleotide polymorphic (SNP) genotypes in a series in inbred strains of mice. There are, however, two important problems, that are to be solved in order to realize this approach: i) how to measure the genotype difference across the strains, and ii) how to access the significance of correlation between the phenotypic and genotypic differences?

Results: We demonstrated that the optimal measure of the genotype difference across the strains is the relative number of distinguishing SNPs in a frame the size of which depends on the map density in various part of genome and include at least 20–30 genotyped positions common to a pair of strains. Substitution of fixed frames by the floating frames led to 3-fold decrease of average size of the genome region where QTL were predicted and increased the number of correctly predicted regions. We suggested and tested method of empiric estimate of the significance of correlation between the phenotypic and genotypic differences based on the resampling realized by a series permutations.

Introduction

A new method for mapping of genes controlling complex traits in mice has been recently suggested by Grupe *et al*. (2001). It was named *in silico* mapping because it does not require generation and analysis of experimental intercross progeny. It is based on computer analysis of databases containing information on the distribution of specific phenotypes and single nucleotide polymorphic (SNP) genotypes in a series of inbred strains of mice. A linkage prediction program scans a murine SNP database and, only on the basis of known inbred strain phenotypes and genotypes, predicts the chromosomal regions that most likely contribute to complex traits. Using the allelic distributions across inbred strains contained in the SNP database, the computational method calculates genotypic distances between regions for a pair of mouse strains. These genotypic distances are then compared with phenotypic differences between the two mouse strains. The process is repeated for all mouse strain pairs for which phenotypic and genotypic information is available. Lastly, a correlation value is derived using linear regression on the phenotypic and genotypic distances for each genomic region.

The rationale of this method appears rather logical and simple. There are, however, two important problems, that have to be solved in order to realize this approach: i) how to measure the genotype difference across the strains, and ii) how to access the significance of correlation between the phenotypic and genotypic differences?

To characterize the genotypic difference between the strains Grupe *et al*. (2001) suggested dividing all chromosomes onto a series of overlapping regions of 30 cM and counting the number of SNP-marker distinguishing the strains located at each of the regions. Smith et al, (2003) proposed a modification of the original method. They suggested calculating a fraction of discriminating

markers present in the regions rather than their absolute number. The size of the regions has been selected arbitrary and based on the density of the SNP map. However, due to the progress in SNP analysis the map becomes the denser and denser day by day. Therefore we need to find a general principle of division of the genome onto series of regions.

To measure a correlation in phenotypic and genotypic differences between the strains Grupe *et al*. (2001) suggested using correlation coefficients calculated for different genome regions. Under the null hypothesis presuming no linkage between the phenotypes and SNP-markers, these coefficients may be considered as independent random values and approximate their distribution by the normal distribution. The author of the method considered a correlation as significant if the standardized correlation coefficient excided 10 % threshold value. The drawback of this approach is that it introduces the same threshold to all regions, although they obviously differ from each other in number and distribution of the SNP-markers and functional significance.

The aims of our study were to analyze the statistical features of the method for *in silico* mapping and their dependence on the measure of genotypic difference and to develop a criteria of significance of contribution of separate genome regions to the control of quantitative trait.

Materials and Methods

Data source. We tested our modification of the method using the data on the phenotypes (number of atherosclerotic plaques) and genotypes for 439942 SNP-markers in mice of 6 inbred strains DBA/2J, C57BL/6, 129/SV-ter, AKR/J, BALB/cByJ, C3H/HeJ (http://aretha.jax.org). The genetic control of this trait has been already analyzed in independent QTL-mapping experiments (http://www.ncbi.nlm.nih.gov/LocusLink/, http://www.informatics.jax.org/). A comparison between the results of the QTL-mapping and *in silico* mapping allowed us to estimate a reliability of the later method

Measure of genotypic differences. The authors of the method suggested dividing genome to a series of overlapping frames and calculating the difference of genotypes for every frame between every pair of strains. The main drawback of this approach is that it uses the frames of fixed size, which differ from each other in the density of markers. To overcome this we suggested to use the frames of floating size, depending in minimal number of typed markers for all pairs of strains. We compared several measures of genotypic differences, varying in the way of calculation (absolute versus relative measure), type of the frame (fixed versus floating) and size of the frame (from 21 to 30 SNPs per frame).

Choice of significant region. To solve the problem of testing of significance of the contribution of the genome region into the control of a trait we suggested determining the empirical threshold level. This approach has been suggested by Churchill and Doerge (1994), it is based on the resampling realized by a series of permutations (Fisher, 1935).

We carried out the resampling as follows. According to a null hypothesis there is no linkage between the marker and the gene controlling the trait. This means that the genotypes and the phenotypes are distributed randomly across the strains. Within the framework of this hypothesis resampling consisted of arbitrary redistribution of the phenotypes across the strains, while their real genotypes remained fixed. We repeated this procedure many times and calculated correlation coefficients between differences of phenotypes and genotypes for every region. Then we selected the threshold value, cutting off a given portion of highest correlations (0.01, 0.05 and 0.10), for every region. If the "real" coefficient exceeds the resampling threshold value at a particular region, then the null hypothesis is rejected for this region, and we may consider the significant contribution of this region into control of the trait.

The main advantage of our method is that it reflects the characteristics of the particular experiment to which it is applied and does not demand any assumptions on the distribution of the quantitative trait. In our study we used 10 % threshold value. Fig. shows the principle of the testing and outlines the significant positions.

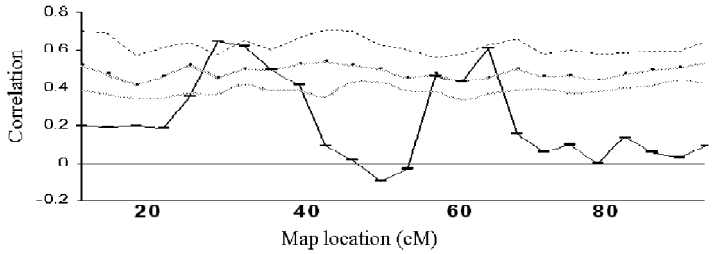


Fig. An example of *in silico* mapping of sample data.
Correlation values coming from the analysis of the empirical data (Smith *et al*., 2003) are connected by solid line. Empirical threshold values (dotted lines) are shown for level $\alpha = 0.01$ (top), 0.05 (middle) and 0.10 (bottom). Arrow indicates a point demonstrated highly significant correlation ($\alpha = 0.01$). Brackets indicate two regions where QTLs are predicted ($\alpha = 0.1$).

Results and Discussion

Table 1 shows the results of testing of six different measures of genotypic differences across the strains. We compared this variants for following indicators: number of *predicted chromosomes* (those for which *in silico* mapping detected regions with a significant contribution into phenotypic difference between the strains); number of *predicted regions*; number of *correctly predicted regions* i.e number of predicted regions matching to those revealed by independent QTL mapping experiment; and *average size of region.*

First of all, we should emphasize, that none of the variant resulted in false chromosome prediction, i.e. indicated the chromosomes other than those twelve detected in independent QTL mapping experiment. A comparison between the absolute and relative measures showed obvious advantage of the latter one (variants 1 vs 2–6, Table). It predicted 10 chromosomes at least, while the former predicted only one chromosome. The relative measure predicted correctly ten regions at least, while the absolute one failed to predict correctly any regions.

Substitution of fixed frames by the floating frames (variants 1–2 vs 3–6, Table) led to 3-fold decrease of average size of the predicted region and increased the number of correctly predicted regions. This indicate that the use of the floating frames increases an accuracy of the prediction.

Decrease of the size of the floating frame (variants 3–6) led to increase of the number of predicted regions. However, the number of new regions (those that have not been detected in independent QTL mapping experiment) increased as well. Unambiguous interpretation of this result is difficult: either *in silico* mapping detected significant regions that slipped detection in QTL mapping, or decrease of the size of the floating frame led to false detection of irrelevant regions due to overestimation of random genetic differences between the strains.

Table. The results of testing of six different measures of genotypic differences across the strains

Variant of analysis	Difference of genotypes	Window's size	Number of predicted chromosomes	Number of predicted regions	Number of correctly predicted regions	Average size of region
1	Absolute	30 cM	1	2	0	16.175
2	Relative	30 cM	10	13	10	16.824
3		21 SNPs	12	23	13	6.403
4		24 SNPs	11	17	12	7.149
5		27 SNPs	11	19	12	8.106
6		30 SNPs	11	15	11	8.847

Thus, our results show that the best measure of genotypic difference is the relative number of distinguishing SNPs in a frame the size of which depends of the map density in various part of genome and include at least 20–30 genotyped positions common to pair of strains.

However, irrespectively to the accuracy of statistical methods, *in silico* mapping has some features that restrict its application. Its results depend not only on the way of measurement of genotypic differences, but also on the accuracy of measurement of phenotypic differences.

Variation of quantitative traits depends on sex, age, rearing conditions of the mice and a number of random environments factors. *In silico* mapping operates with between-strain difference in the trait and ignores within-strain difference. Therefore a bias in estimate of averages may led to distortion of the mapping results. This method assumes additive gene action and therefore ignores within- and between-region interactions. Thus, this method may be recommended as a tool for fast preliminary analysis of linkage, while the final mapping should be based of the analysis of crosses producing genetically heterogeneous progeny.

Acknowledgements

The work was supported by the grants RFBR (040448024 and 040448074) and Programs of Russian Academy of Sciences (24 and 25).

References

Churchill G.A., Doerge R.W. Empirical threshold values for quantitative trait mapping // Genetics. 1994. V. 138. P. 963–971.

Fisher R.A. The design of experiments. Ed.3. Oliver & Boyd Ltd., London. 1935.

Grupe A., Germer S., Usuka J., Aud D., Belknap J.K., Klein R.F., Ahluwalia M.K., Higuchi R., Peltz G. *In silico* mapping of complex disease-related traits in mice // Science. 2001. V. 292. P. 1915–1918.

Smith J.D., James D., Dansky H.M., Wittkowski K.M., Moore K.J., Breslow J.L. *In silico* quantitative trait locus map for atherosclerosis susceptibility in apolipoprotein E-deficient mice // Arterioscler Thromb Vasc Biol. 2003. V. 23. P. 117–122.

BGRS
2004

COMPUTATIONAL STRUCTURAL AND FUNCTIONAL PROTEOMICS

PREDICTING CONTACT NUMBERS OF AMINO ACID RESIDUES USING A NEURAL NETWORK MODEL

Afonnikov D.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; Novosibirsk State University, Novosibirsk, Russia, e-mail: ada@bionet.nsc.ru

Keywords: *protein structure, residue contact numbers, neural network*

Summary

Motivation: The number of inter-residue contacts is an important structural characteristic of residues in protein. The characteristic is related to solvent accessibility of residue, and it may be used for predicting the protein contact maps.

Results: Here, an approach to the prediction of the number of inter-residue contacts, based on protein primary structure using the physicochemical properties of amino acids, and a neural network algorithm is proposed.

Introduction

In the post-genomic era, a challenge is development of algorithms and programs for predicting the structural/functional properties of proteins. The goal is to achieve the most complete annotation of the protein moiety of the newly sequenced and well-characterized genomes on the basis of their sequences. One of the prediction categories for protein structure is the number of inter-residue contacts (Fariselli, Casadio, 2000). The approaches to estimation of the number of residue contacts in proteins rely on neural network algorithms (Fariselli, Casadio, 2000; Pollastri *et al*., 2001), with one of the two states of a given residue predicted, depending on whether the number of contacts is lower or higher than its average distribution value. However, there are tasks when little information can be derived from the model of the two state predictions. For example, the residue contact value can be used as an additional parameter for an efficient search of the native contact map (Kabakcioglu *et al.*, 2002). With these considerations, the development of methods giving a real contact number for a residue appears timely. It should be noted that, in the case of prediction of solvent accessibility, a method for calculating the real accessibility number has been proposed (Ahmad *et al.*, 2003).

The method suggested here is based on a neural network allowing estimating the real contact number of a residue at a given protein position. The task of predicting the number of inter-residue local contacts (with the chain neighboring residues) is also considered. The prediction accuracy for the number of total and local contacts is compared.

Methods

Design and training of neural networks. For the i-th position in protein primary structure, a window of the 2h+1 size is considered; the i residue is in the center of the window and its boundaries cover residues with numbers the i-h, i+h (here, h=8). Prediction is made for the i-th residue based on information on the physicochemical properties of amino acids in the window. We considered 5 physicochemical properties: volume, polarity, isoelectric point, hydrophobicity and water accessibility (Bogardt *et al.*, 1980). The terminal residues within the h positions from the N and C termini of a polypeptide chain were excluded from consideration. The thus calculated numerical parameters were input into the neural network, which estimated the contact number for the i-th residue. A network consisting of two internal layers with 5 and 9 neurons each was chosen in such a way that prediction was best without overtraining effect. The neural network was trained and

tested using a sample of 721 monomeric proteins with known spatial structures in the PDB databank (Berman *et al.*, 2000). The protein sequences had pairwise similarity not exceeding 40 %, the resolution of the structures was greater than 2 Å, they did not contain missed CA-atom coordinates. The sample was divided into a training set of 625 proteins and a testing set of 156 proteins with about the same distribution along the chains. The distance between two CA-atoms, which did not exceed r=10 Å, was accepted as the contact distance. For amino acids of different types at the central position of the window, the neural network was trained separately. As a result, 20 neural networks were generated. Their topology was the same; their weights were different, depending on the type of the central residues.

Prediction of the numbers of local contacts. It should be noted that the total number of contacts n_{tc} is the sum of the contacts considered with respect to the distance along the protein chain: the nearest neighbors n_{lc} (the number of local contacts) and the furthest neighbors. The n_{lc} parameter can strongly depend on the secondary structure of the polypeptide chain and it can be largely due to local residue interactions. In the current work, 20 neural networks corresponding to the number of amino acid types to predict the number of local contacts were built. These networks had the same topology and input parameter set as described above. The output parameter for these networks was the number of contacts with the residues at a sequence distance not greater than h from the i-th residue.

Prediction of the numbers of local contacts. It should be noted that the total number of contacts n_{tc} is the sum of the contacts considered with respect to the distance along the protein chain: the nearest neighbors n_{lc} (the number of local contacts) and the furthest neighbors. The n_{lc} parameter can strongly depend on the secondary structure of the polypeptide chain and it can be largely due to local residue interactions. In the current work, 20 neural networks corresponding to the number of amino acid types to predict the number of local contacts were built. These networks had the same topology and input parameter set as described above. The output parameter for these networks was the number of contacts with the residues at a sequence distance not greater than h from the i-th residue.

Prediction accuracy. Let us denote by n and n', respectively, the real and predicted total contact number of residues in the testing sample of size N. Then, $dn = \frac{1}{N}\sum_{k=1}^{N}|n_k - n_k'|$ will stand for the absolute value of the deviation of the predicted contact number from the real, and the $rn = dn/(\frac{1}{N}\sum_{k=1}^{N} n_k)$ parameter will express its ratio to the average of the real contact number. The smaller the two parameters are, the more accurate are the predictions of the contact number for the residues.

The Q(x) parameter denotes the fraction of the predicted contact numbers which differ from the real by a value not exceeding *x*. The greater is the Q(x) parameter, the more accurate is the prediction. For comparison, x=0 (accurate prediction) and *x*=2 (prediction with an absolute error not greater than 2) were taken. These parameters were calculated for predicting the number of both total and local contacts.

Baseline algorithm (BAL). The accuracy of prediction by the neural network algorithm was compared with the one of the BAL, which disregards the residue local environment, being dependent on the choice of the most frequent contact number for an amino acid.

Results and Discussion

The results for the prediction accuracy of the numbers of the inter-residue contacts are set out in Table. In the case of prediction of the total number of contacts, the accuracy of the neural network is slightly greater than that of the BAL. In the case of the prediction of the number of local contacts, the neural network algorithm allowed to achieve an increase in accuracy prediction by

6 % for Q(2) and by 17 % for *dn* in comparison with the BAL. This result is much better than for the number of total contacts; the *rn* error value for the local contacts is also the smallest. The improvement may be due to the greater dependency of the local contact number than of the total on the local interactions with the nearest neighbors. Therefore, the local contacts can be more accurately calculated using only the local sequence context.

Table. A comparison of the prediction accuracy for the number of inter-residue contacts obtained by the baseline (BAL) and the neural network (NN) algorithm. The estimates are for the total contact (n_{tc}) and local contact (n_{lc}) numbers

Contact number	Method	Q(0)	Q(2)	dn	rn
n_{fc}	BAL	7.25%	34.90%	4.63	0.24
	NN	7.37%	35.50%	4.17	0.22
n_{nc}	BAL	15.02%	62.13%	2.26	0.23
	NN	14.81%	68.00%	1.93	0.19

Prediction accuracy of the contact numbers is also dependent on the type of the central residue. Figure presents the dn values for 20 amino acid types for the n_{tc} and n_{lc} parameters and also for two prediction types.

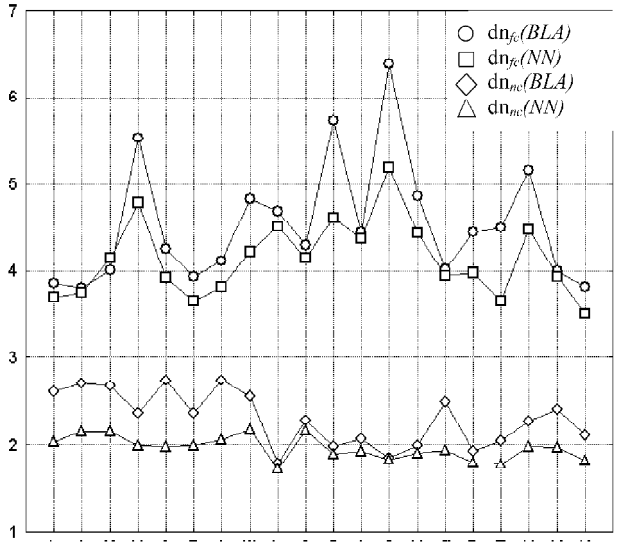


Fig. The dn values for 20 amino acids designated by capital letters under the X-axis predicted by two different algorithms and contact number types. The designations of the curves are at the right.

It is important to note that many factors (the occurrence of residues of a particular type, the physicochemical property of a given residue, its propensity to form a particular secondary structure, among others) can affect prediction accuracy. For example, because of the specific structure of a side chain, covalently bound to the backbone, proline occurs more frequently in loop conformations; as a result, prediction accuracy is affected so that the number of predicted local contacts, strongly dependent on the secondary structure for praline, is just as well predicted by BAL and the neural network. This means that the local structure of the polypeptide chain is weakly dependent on the nearest neighbor residues, and it is almost entirely dependent on the structural features of proline itself. Similar results were observed for glycine, whose backbone is very flexible because it has no side chain and occurs more frequently in the loop conformation. Taken together, evidence was obtained for the higher prediction accuracy of the local inter-residue contacts by the neural network predictor.

Acknowledgements

This work was partly supported by grants from the Russian Foundation for Basic Research (03-07-96833-p2003,03-07-96833); Siberian Branch of the Russian Academy of Sciences (projects Nos. 148, 119); the Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501, subcontract 38/2004), RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4), the CRDF and the Ministry of Education of Russian Federation within the Basic Research and Higher Education Program (Y1-B-08-20).

References

Ahmad S., Gromiha M.M., Sarai A. Real value prediction of solvent accessibility from amino acid sequence // Proteins. 2003. V. 50. P. 629–635.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank // Nucl. Acids Res. 2000. V. 28. P. 235–242.

Bogardt R.A. Jr., Jones B.N., Dwulet F.E., Garner W.H., Lehman L.D., Gurd F.R. Evolution of the amino acid substitution in the mammalian myoglobin gene // J. Mol. Evol. 1980. V. 15. P. 197–218.

Fariselli P., Casadio R. Prediction of the number of residue contacts in proteins // Proc. Int. Conf. Intell. Syst. Mol. Biol. 2000. V. 8. P. 146–151.

Kabakcioglu A., Kanter I., Vendruscolo M., Domany E. Statistical properties of contact vectors // Phys. Rev. E Soft. Matter. Phys. 2002. V. 65(4 Pt 1). 041904.

Pollastri G., Baldi P., Fariselli P., Casadio R. Improved prediction of the number of residue contacts in proteins by recurrent neural networks // Bioinformatics. 2001. V. 17. Suppl 1. S234–242.

COMBINED APPROACH TO PROTEIN SECONDARY STRUCTURE PREDICTION

Amirova S.R.*, Machavariani M.A., Filatov I.V., Milchevsky Ju.V., Esipova N.G., Tumanyan V.G.

Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
***** Corresponding author: e-mail: amirova_s@hotmail.com

Keywords: *protein secondary structure, rotamers library, conformational calculations, discriminant analysis*

Summary

Motivation: The most important achievements in protein secondary structure prediction are based on two different approaches. The first one is the statistical approach and the second one is the physical-chemical approach. In the first approach we analyze appearance of different types of amino acids in given conformations. The second approach use conformational calculations and physical-chemical properties of a given molecule. Presently these approaches are developed independently. The creation of the method, which will contain advantages of the statistical and the physical-chemical approaches, is very important task.

Results: We have developed the new approach for secondary structure prediction. Using our combined approach one can obtain the secondary structure of a given protein from its primary structure only. The base of our method is a joint using advantages provided by conformational calculations, data on primary structure and physical-chemical properties of proteins. For the combined approach to be demonstrated, we have predicted the protein secondary structure of four basic types: α-helix, helix 3/10, coil, and turn using only sequences of given proteins from Protein Data Bank.

Introduction

Presently is reached the essential progress in the protein spatial structure prediction (Bonneau, Baker, 2001; Lim, 1974). The most important purposes of current researches are the fundamental understanding of physical-chemical principles, which define stability of a given protein. For a spatial model of protein to be created, it is necessary to take into account internal and external molecular interactions in a given protein. Molecular mechanical and molecular dynamical methods are widely used for these tasks. These methods allow to reproduce the structure of a given molecule by means of the certain force field and the finding of the global energy minimum (Milchevsky *et al.*, 2001).

The base of our method is a joint using results provided by conformational calculations, data on primary structure and physical-chemical properties of model protein structures. These results are necessary for further statistical calculation by stepwise discriminant analysis.

Discriminant analysis is used with the aim to find discriminant functions. These functions divide (or discrimine) two or more category to the best advantage. Also one can use discriminant analysis with the aim to find variables, which provide the statistical significant contribution in the division on categories.

Model

Our method has been developed specially for simultaneous use of molecular-mechanics calculations, data on primary structure and the physical-chemical properties of proteins. Conformational computations for polypeptide chains of some proteins are carried out for different backbone

conformations. These conformations correspond to basic types of secondary structure. Optimization procedure is implemented using the rotamers library (Dunbrack, Cohen, 1997) with the aim to find optimal conformations of side chains. The best values of observed energy and its components are a subset of predictor variables for further statistical analysis by stepwise discriminant analysis. All variables with the exclusion of predictor variables represent sequence and physical-chemical properties of proteins. During discriminant analysis we predict classification into a given category as a function of the predictor variables. Developed program enables to handle more then 1000 given categories and 10000 predictor variables. The preliminary analysis of Protein Data Bank has shown that a bigger part of real peptides are contained in the smaller amount of categories. The method investigates the statistical significance for every predictor. As a result, statistically insignificant variables not influence on category classification. The method provides quantitative measure for comparison of models based on different sets of predictor variables. The final purpose of calculations is the finding discriminant functions. Then on their base we obtain the probability for every amino acid to be in the conformation of four basics secondary structure types.

From Protein Data Bank we took 48 proteins with the following codes:
1A6N,1aon,1AYE,1baz,1bni,1BRS,1c8c,1c9o,1cei,1CSE,1CSP,1divm,1eal,fkb,1FNF,1g6p,1HNG, 1hz6,1IMQ,1lmb,1LOP,1mjc,1OPA,1pgb,1PHP,1pin,1PNJ,1poh,1psf,1QOP,1RA9,1ris,1shf, 1SHG,1srl,1ten,1TIT,1urn,1wit,256b,2A5E,2ABD,2CI2,2LZM2PDD,2RN2,2VIK,3CHY

For each primary structure corresponding to real protein we have executed starting-up procedure with the aim to get the initial backbone conformation. To find the optimal conformation we were varying dihedral angles of side chain in accordance with the limited selection algorithm. Then we optimized obtained structures using secondary structure dependent rotamer library (SSDEP). After what, we have done energy optimisation using Internal Coordinates Mechanical (ICM) method (Abagyan *et al*., 1994). From all obtained structures we have chosen the structure with the best conformational energy and considered it as final result.

Thereby we have four categories in accordance with four basics secondary structures types. As for input variables, we have chosen all energy compounds, which we used at minimization: Van-der-Waals energy, hydrogen bonding energy, torsion energy, electrostatic energy and full conformational energy. For discriminant functions to be sensitive to not only linear dependences from energy, we included in the set of input variables squares and cubes of all used energy terms. Finally we had fifteen input variables. The most statistically important variable was the square of Van-der-Waals energy term.

Results and Discussion

After discriminant functions are known our program calculates the probability for every amino acid to be in the conformation of four basics secondary structure types. These probabilities for certain amino acids from a several of 48 accepted in attention real proteins to be in conformations of four basics secondary structure types are shown in Table 1. So Table 1 consists only the part of our results.

In our method the amino acid is distributed correctly on secondary structure type, if it has the maximal probability to be in the conformation of the certain secondary structure type in comparison with probabilities to be in conformations of the others secondary structure types. Using such approach of distribution we obtain the Table 2, which gives results of protein secondary structure prediction.
The every line in Table 2 corresponds to the certain secondary structure type in a given protein. For example considering the secondary structure type α-helix in all accepted in attention proteins with known secondary structure we have 1125 + 693 + 351 + 225 = 2394 amino acids in α-helix conformation. According to our calculations 1125 of them are in α-helix conformation, 693 are in helix 3/10 conformation, 351 are in coil conformation and 225 are in turn conformation.

Table 1. Probabilities for amino acids to be in conformations of four basics secondary structure types

The central amino acid in pentapeptid from the secondary structure of a real protein	The probability for the amino acid to be in the conformation of the certain secondary structure type			
	AlphaHelix	Helix 3/10	Coil	Turn
AlphaHelix, GLU	0.012	0.975	0.003	0.007
AlphaHelix, ALA	0.613	0.170	0.014	0.200
AlphaHelix, MET	0.066	0.871	0.008	0.053
AlphaHelix, ASP	0.228	0.560	0.054	0.156
AlphaHelix, LYS	0.414	0.336	0.051	0.196
AlphaHelix, VAL	0.276	0.133	0.123	0.466
Turn, GLY	0.426	0.019	0.119	0.434
Turn, LYS	0.323	0.019	0.230	0.426
Turn, GLU	0.207	0.004	0.397	0.390
Turn, GLY	0.265	0.000	0.145	0.589
Coil, GLY	0.096	0.287	0.184	0.430
Coil, THR	0.030	0.725	0.131	0.112

Table 2. Results of protein secondary structure prediction

	α-helix	helix 3/10	coil	turn
α-helix	1125	693	351	225
helix 3/10	1	42	2	9
Coil	288	387	711	117
Turn	198	324	117	90

From the Table 2 one can evaluate the accuracy of our method using the percent of correct prediction secondary structure type. For helix 76 % of amino acids in all accepted in attention proteins were predicted correctly, for coil the percent of corrected prediction is 47 %, and for turn the percent is– 12 %.

The main limitation in the application of our method is an ambiguity in the classification on secondary structure types for a given backbone, because there are many different classifications on secondary structure types presently.

Using the Table 2 one can see the main advantages and disadvantages of our method. One of the most important advantages is the possibility of simultaneous prediction of several secondary structure types being kept in a given sequence.

As for disadvantages, one of them is the comparatively low percent of correct prediction secondary structure type presently.

The main trend of the improvement of the method is the using more detailed definitions of the secondary structure types and the assignment larger number of categories for them. For example using the DSSP classification (Kabsch, Sander, 1983) one can find a several groups of ten of b-turns. In this classification many conformations, which formally correspond to different types of b-turns, have absolutely different set of backbone dihedral angles. Preliminary it is necessary to distribute different secondary structure types on intelligent subcategories. After distribution we can attribute several secondary structure types in one category coming from physical considerations.

Acknowledgements

This work was supported by grants from the Russian Foundation for Basic Research (No. 03-04-49017 and No. 02-04-49114), grant for Ministry of Industry, Science and Technologies (No. 43.071.1.1.1517) and grant on Molecular and Cellular Biology RAS (Program No. 10).

References

Abagyan R.A., Totrov M.M., Kuznetsov D.N. ICM – a new method for protein modeling and design. Applications to docking and structure prediction from the distorted native conformation // J. Comp. Chem. 1994. V. 15. P. 488–506.

Bonneau R., Baker D. Ab initio protein structure prediction: progress and prospects // Annu. Rev. Biophys. Biomol. Struct. 2001. V. 30. P. 173–189.

Dunbrack R.L., Cohen F.E. Bayesian statistical analysis of protein side-chain rotamer preferences // Protein Sci. 1997. V. 6. P. 1661–1681.

Kabsch W., Sander Ch. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features // Biopolymers. 1983. V. 22. P. 2577–2637.

Lim V.I. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins // J. Mol. Biol. 1974. V. 88. P. 873–894.

Milchevsky J.V., Esipova N.G., Tumanyan V.G. *et al*. Molecular modelling of disease-causing single-nucleotide polymorphisms in collagen // SAR QSAR Environ Res. 2001. V.12(4). P. 383–99.

STRUCTURE AND POLYMORPHISM OF THE HIV-1 PRINCIPAL NEUTRALIZING EPITOPE

Andrianov A.M.*

Institute of Bioorganic Chemistry of National Academy of Sciences, Minsk, Republic of Belarus
* Corresponding author: e-mail: andrianov@iboch.bas-net.by

Keywords: *human immunodeficiency virus, protein gp120, principal neutralizing epitope, three-dimensional structure, computer modeling, NMR spectroscopy*

Summary

Motivation: The high degree of conservation of certain amino acid positions residing in the immunogenic tip of the HIV-1 gp120 V3 loop implies that the residues occupying them have a key structural and functional role and are important in some aspects of virus infectivity, such as modulation of the conformation of the V3 loop or modulation of the interaction with coreceptor. Therefore, knowing the structure of the V3 loop principal neutralization site is essential for understanding its effect on viral tropism.

Results: The high resolution 3D structure models were constructed for the HIV-Haiti and HIV-RF immunogenic crowns, and their geometric parameters were collated with the ones of conformers derived previously for describing the conformational features of immunogenic tip of gp120 from Thailand and MN HIV-1 strains. The HIV-1 principal neutralization site was demonstrated to make in water solution highly flexible system sensitive to its environment. This inference is completely valid for the geometric space of dihedral angles where statistically significant differences in local structures of simulated conformers have been found for all virus isolates of interest. In spite of this fact the stretch analyzed was shown to manifest certain conservatism in the space of atomic coordinates, building up in four HIV-1 isolates two spatial folds close to those observed in crystal for the V3 loop peptides bound to different neutralizing Fabs.

Introduction

The HIV-1 principal neutralizing determinant is located within the third variable loop (V3) of envelope protein gp120 with disulfide-bridged invariant cysteines. This loop plays a critical role in determining chemokine receptor usage, and it not only determines chemokine receptor specificity but also helps immune system evasion. The V3 loops of different HIV-1 isolates contain highly variable amino acid sequences, which prevents antibodies against a V3 loop, bound to one isolate, from having effect on the V3 loop of a different isolate. Therefore, for a deeper understanding of the nature of specific interactions giving rise to a viral antigen-antibody complex and of the role of conserved regions in this process, one must know the structure of gp120 neutralizing epitope in different HIV-1 strains. The comprehensive analysis of literature data on the structure of HIV-1 immunogenic crown indicates that they differ significantly among themselves and give discrepant information, which needs more detailed studies.

This report continues our theoretical studies (Andrianov, 2002; Andrianov, Sokolov, 2003; 2004) in which the evidence for the local structures as well as for the most probable spatial folds of Thailand and MN HIV-1 immunogenic tip were derived on the basis of NMR spectroscopy data. Here, we built the NMR-based 3D structure models for the immunogenic tip of Haiti and RF HIV-isolates (hexapeptide fragment located within the central region of the V3 loop) and compared them with those obtained previously (Andrianov, 2002; Andrianov, Sokolov, 2003; 2004) for Thailand and MN virus strains. To this end, (i) the V3 loop fragments including the stretches in question were subjected to computerized NMR analysis to establish all of the conformers consistent

with the experimental data and energy criteria; (ii) the best energy hexapeptide structures derived from these computations were exposed to quantum chemical refinement to estimate their relative populations in the simulated ensembles; (iii) the prevalent structures thus obtained were collated with the X-ray conformations of the homologous stretches of HIV-MN V3 loop peptides; (iv) on this basis, the 3D structure model was constructed for the HIV-1 principal neutralization site.

Model

A hierarchical procedure (Andrianov, Sokolov, 2003), using a "bottom-up" strategy and combining (i) a probabilistic approach for estimating all possible starting structures, (ii) restrained molecular mechanics algorithms for preliminary selection of all energetically preferred conformers, as well as (iii) quantum chemical computations for refining their geometry, was called for action to study the structural properties of HIV-Haiti and HIV-RF neutralizing epitope in terms of NMR spectroscopy data.

At the first point of this procedure the NMR spectral parameters are directly (without building a 3D structure) transformed into the dihedral angles of the amino acid residues, which are considered as the initial approximation for the search in the conformational space of a full distribution of structures consistent with theoretical data and experimental observations. The residue conformations of the V3 loop peptides were determined using a probabilistic model of protein conformation and calculating the mean weighted values of dihedral angles from the NMR data (Catasti P. et al., 1996), which were analyzed statistically with account for the empirical distribution function for main chain torsional angles built on high-resolution X-ray data. The calculations were performed with the CONFNMR-2 program (Andrianov, 2002) that used the data on d-connectivities (Catasti *et al*., 1996) to determine the conformational space (ϕ,ψ) region for every amino acid residue and then calculated the most probable dihedral angles and corresponding standard deviations.

The conformational analysis (step 2) was performed by the conventional molecular mechanics methods assuming the standard residue geometry. When calculating intramolecular energy, non-bonded and electrostatical interactions as well as torsional potentials and hydrogen bonds were taken into account. The criterion for selecting conformationally stable structures was an energy interval $\Delta U = 11$ kcal/mole.

Refining the structures as well as determining the formation heats were carried out by the PM3 quantum chemical method. The mean weighted ϕ,ψ angles for conformers selected were used as the initial approximations for building the final structures. To attain an absolute consent between geometry of these structures and NMR requirements, use was made of theoretical approach (Andrianov, Sokolov, 2003) which included the trial computational experiments to find an optimum conditions of geometry optimization.

Results and Discussion

The values of dihedral angles and interatomic distances in three conformers which were selected for the HIV-Haiti immunogenic crown display that its backbone can form the different secondary structure elements in water solution. So the geometric parameters observed in conformer 1 match a double β-turn IV-IV (residues 1–4 and 3–6). In conformers 2 and 3 they fit into a triple β-turn IV-IV-IV (residues 1–4, 2–5 and 3–6) as well as into a single β-turn IV (residues 1–4) respectively. (Here and further the simulated conformers are numerated in order of increasing the formation heat values). These differences in local conformer structures give rise to the different spatial hexapeptide shapes whose specific feature is a compact fold of its polypeptide chain.

The pairwise best matching the conformers simulated here to the X-ray structures of the V3 loop synthetic molecules (Stanfield, 1999) implies that the overall shape that is realized in conformer 1 is just like the one revealed in crystal for homologous site of peptides bound to the Fab fragments of

antibodies 50.1, 59.1, and 83.1. At the same time the spatial hexapeptide forms characteristic of conformer 3 as well as of three molecules bound to the Fab fragment of antibody 58.2 are also alike.

In such a manner, two structures of the HIV-1 immunogenic crown found in crystal (Stanfield, 1999) are also present in the list of preferred conformers for the immunogenic crown of Haiti-V3 loop peptide. This conclusion has a bearing upon the 3D hexapeptide forms which are composed of different local minima of its amino acid residues: collating the simulated and crystal conformations in the geometric dihedral space testifies to the fact that in both cases under study the differences between them are statistically significant.

The maximum energy barriers between conformers 1, 2 and 2, 3 equal to 6.6 kcal/mole are sufficiently large, whereas the corresponding value for structures 1 and 3 equal to 3.6 kcal/mole indicates that the structural transition occurring in crystal upon substituting the antibodies (Stanfield, 1999) may also take place in solution upon varying the hexapeptide environment. Since the energy barriers between conformers 1 and 3 are unimportant for the stretch in question, the alterations of its environment caused by replacing the antibody may lead to the redistribution of conformer statistical weights shifting the maximum of the occupancies of the conformational states towards one of them ensuring the virus interaction with a given antibody. These evidence pointing out the probable conformational mimicry of gp120 stretch analyzed are in line with the X-ray data (Stanfield, 1999) and substantiate earlier supposition (Andrianov, 2002) according to which its peptide chain exhibits in water solution the cluster of conformers and only one of them, most likely a dominant structure, determines the specificity of virus binding with antibodies. In other words, in particular case of Haiti-V3 loop a double β-turn structure prevalent in the ensemble of preferred conformers in the unbound state plays the role of "signal" conformation which is recognized by the immune system.

Having analyzed the geometric parameters for the best energy conformer derived for the HIV-RF immunogenic tip, I surprisingly revealed that its secondary structure is identical to that found earlier (Andrianov, Sokolov, 2003) for the prevalent conformation of the principal neutralizing epitope from Thailand HIV-1 isolate. This structural element is composed of the inverse g-turn which resides in the invariant crest Gly-Pro-Gly as well as of the non-standard b-turn IV located at the overlapping tetrapeptide Gly-Arg-Val-Ile.

Thus, according to the evidence above the conformation of the inverse g-turn that is infrequent in proteins manifests at the Gly-Pro-Gly site in the dominant structures of the V3 loop peptides derived from Thailand and RF HIV-1 strains. Reasoning from this fact, I believe that in particular cases in question this structural element may play an important role in the process of virus binding with antibodies.

The results of superimposition of conformers 1 on the one hand as well as of the X-ray structures for homologous sites of the V3 loop peptides in complexes with different neutralizing Fabs (Stanfield, 1999) on the other hand illustrate the close agreement of NMR-based structure 1 with the crystal conformation of synthetic molecules bound to the Fab fragment of antibody 58.2. As with the HIV-Haiti isolate, this consensus deals with the 3D hexapeptide folds and does not concern its local conformations which have the great discords among the values of dihedral ϕ,ψ angles in structures compared. These data confirm the inference made in preceding study (Andrianov, 2002) where I state that the "basic" Gly-Pro-Gly-Arg-Ala-Phe sequence and fragments homologous to it have a large reserve stock of the "conformational strength" that promotes for them to preserve the close spatial folds in substantially different force fields.

Comprehensive analysis of the all data obtained allows to suppose that the structural transition turning up in crystal as a result of replacing the antibodies (Stanfield, 1999) may also occur in the unbound state in going from one to another HIV-1 isolate: in this case, the modification of hexapeptide environment required for overcoming the energy barriers between conformers can be

induced by the high mutability of the V3 loop amino acid sequence. Positively, collating the ensembles of conformers computed here and previously (Andrianov, Sokolov, 2003; Andrianov, Sokolov, 2004) substantiates this supposition: according to these calculations, a double b-turn found as the best energy structure for the immunogenic crown of Haiti and MN isolates is present as a minor conformer in the list of the putative conformations for homologous site of Thailand HIV-1 strain (Andrianov, Sokolov, 2003). Moreover, the structure prevailing in the Thailand and RF isolates has been revealed as a marginal hexapeptide conformation in the HIV-MN isolate (Andrianov, Sokolov, 2004).

In such a manner, the data obtained prove convincingly that the HIV-1 principal neutralizing epitope exhibits in water solution the flexible system sensitive to its environment.

Acknowledgements

This work was supported by grant from the Byelorussian Foundation of Fundamental Investigations (X04-058).

References

Andrianov A.M. Local structural properties of the V3 loop of Thailand HIV-1 isolate // J. Biomol. Struct. Dynam. 2002. V. 19. P. 973–990.

Andrianov A.M., Sokolov Yu.A. Structure and polymorphism of the principal neutralization site of Thailand HIV-1 isolate // J. Biomol. Struct. Dynam. 2003. V. 20. P. 603–614.

Andrianov A.M., Sokolov Yu.A. 3D structure model of the principal neutralizing epitope of Minnesota HIV-1 isolate // J. Biomol. Struct. Dynam. 2004. V. 21. P. 577–590.

Catasti P., Bradbury E.M., Gupta G. Structure and polymorphism of HIV-1 third variable loops // J. Biol. Chem. 1996. V. 271. P. 8236–8242.

Stanfield R.L., Cabezas E., Satterthwait A.C., Stura E.A., Profy A.T., Wilson I.A. Dual conformations for the HIV-1 gp120 V3 loop in complexes with different neutralizing Fabs // Structure. 1999. V. 7. P. 131–142.

NEW APPROACHES TO ANALYSIS OF PROTEIN STRUCTURE AND FUNCTION

Bachinsky A.G.*, Solovyev V.V.

Softberry Inc., 16 Radio Circle, Suite 400, Mount Kisco, NY 10549, USA
***** Corresponding author, Permanent address: SRC VB "Vector", Koltsovo, Novosibirsk region, 630559, Russia; e-mail: bachin@vector.nsc.ru

Keywords: *protein structure and function, evolution, computer analysis, disordered regions, subcellular locations*

Summary

Motivation: Prediction of protein function and structure using computational tools acquires more importance as the gap between the increasing amount of protein sequences and their experimental characterization widens. The computational approaches allow revealing new and elaborate known evolutionary relations between proteins, and help to determine the role of protein in metabolism, its interaction with the other participants of cellular processes.

Results: New approaches are presented in this report for analysis of the structure and function of proteins, such as predicting secondary structure and disordered regions, fold recognition, determination of protein sub-cellular location and identification of cysteine forming SS-bonds.

Availability: http://www.softberry.com

Programs:

1. Prediction of protein secondary structure using local alignments ('alpha', 'beta', 'coil'), the SSPAL program

We have developed a further variant of the nearest-neighbor approach, which uses local alignments of query sequence with a non-redundant set of PDB sequences (Salamov, Solovyev, 1997). Alignment score depends not only on amino acids per se, but also takes into account similarity of the secondary structure and preference to be buried or accessible to polar solvent. The average accuracy of prediction is ~ 70 %. It increases if we provide the program with the multiple alignment generated from sequences selected from the NR database similar with the query sequence and compare this alignment not with the PDB sequences themselves, but with multiple alignments constructed in advance for each PDB sequence.

2. Alignment of query sequences with sequences of the PDB databank and their fold identification, the FOLD program

A non-redundant database of 3D-structures corresponding to the SCOP protein domains has been generated. The FOLD program uses comparison of query sequence and its predicted secondary structure with the SCOP domains. For each pair of query – database sequences, a dot-matrix is built to compute sequence structure alignment. The dot-matrix score is a combination of the sequence similarity score, the secondary structure similarity score and the environment preference in a window of length 9. It is possible to compute the global or local alignment. We found that local alignments are better than global corresponding to structured alignments generated by CE program (Shindyalov, Bourne, 1998). A set of accompanying programs allows comparing alignments of a query sequence with representatives of different SCOP domains, as well as to get coverings of the query by an optimum set of local alignments with proteins of known 3D structure.

3. Prediction of disordered regions of proteins, the PDISORDER program

The current dogma holds that an amino acid sequence determines the 3-D protein structure and

the resulting 3-D is crucial for protein function. However, *many proteins (or their fragments) remain unfolded under physiological conditions, yet carry out function* (Li *et al*., 2000). It is hypothesized that, since an amino acid sequence determines the 3-D structure, the sequence should determine the lack of the 3-D structure as well. If this is true, the accuracies of disorder predictions using amino acid sequence information would exceed the accuracies expected by chance (Dunker, Obradovic, 2002).

Training: All disordered regions data used in this work have been downloaded from http://disorder.chem.wsu.edu (649 sequences, 61237 disordered positions). Ordered data were generated from fragments of high resolution 3D structures that were selected from nonredundant PDB set (2,017 fragments, 309,454 ordered positions). Different composition-based attributes (Li *et al*., 2000), as well as numerus property-based attributes (including all the features from the Aaindex: http://www.genome.ad.jp), were tested of their significance of discrimination by the Linear Discriminant Analysis. Two sets of significant attributes: one for the Neural Network, the other for the Linear Discriminant Function were selected using the analytic LDA procedure. Several approaches were tested for recognition of disordered and ordered positions in proteins. A combination of the Neural Network, the Linear Discriminant Function and the Smoothing Procedure was finally constructed. Three windowing procedures were used, called the left, right and intermediate. For all the windows, the attributes are calculated through 31 residues.

Tests: Performance of the PDISORDER program ***was compared with the PONDR*** **and GlobPlot**. Programs. It exceeds them and achieves ~90 %.

4. Recognition of cysteines, forming SS-bonds, the CYS_REC program

For recognition of the SS-bound cysteines in amino acid sequences, the following study was performed.

Training: Coordinates of sulfur atoms were extracted from the PDB entries. The distances between sulfur atoms were calculated for the PDB entries with a resolution not less than 2E. If the distance was less or 2.1E, it was considered that the CYS amino acids form a SS-bond. If the distance exceeded 6E, then the CYS molecule was considered as 'unbound'. Cysteines with the distances in the interval 2.1E – 6E did not participate in the training. Sequences and predicted secondary structures in the ±10 positions interval from the "bound" and "unbound" cysteines formed "positive" and "negative" samples. Using these samples, we computed the "weight matrices" based on the frequencies of amino acids and secondary structure types at the corresponding positions of the selected fragments (Gribskov *et al*., 1987), the log-odds BLOSUM-like matrix of interchangeabilities of amino acids and secondary structures (Henikoff S., Henikoff J., 1993). A linear discriminant function was developed that was based on four features, the most significant for the division of the positive and negative examples.

Prediction: The secondary structure is predicted for a query sequence. A fragment in ±10 positions interval from any cysteine is compared with such fragments of training sets using the log-odds matrix, and the maximum score is defined for each set. The scores of comparisons with profiles of positive and negative examples are calculated for a given fragment, as well as the value of the linear discriminant function. The resulting score computed as a linear combination of five scores listed above is used for the recognition.

Tests: Approximately 3.000 positive and 3.000 negative examples were prepared from the PDB sequences not involved in the training. An accuracy of recognition by combined function on this control set was ~90 %, while the accuracy provided by each component separately did not exceed 75 %.

5. Program for Identification of sub-cellular localization of proteins: ProtComp

ProtComp combines several methods to predict protein localization: neural networks-based prediction; prediction based on linear discriminant analysis; direct comparison with database of proteins with known location; search for certain location-specific motifs; prediction of certain

functional peptide sequences (such as signal peptides, signal-anchors, GPI-anchors, transit peptides of mitochondria and chloroplasts and transmembrane segments for eukaryotic proteins. The programs include separately trained recognizers for proteins of animals+fungi (9 localizations), plants (8 localizations) and bacteria (4 localizations). The following table provides estimates of the prediction accuracy for each compartment of an animals+fungi set of proteins.

Compartment	Sample Size	Accuracy estimates, %
Nucleus	200	88
Plasma Membrane	200	87
Extracellular	162	83
Cytoplasm	200	63
Mitochondria	85	82
Endoplasmic Reticulum	46	83
Peroxisome	35	97
Lysosome	23	91
Golgi	26	77

References

Dunker A.K., Obradovic Z. The Protein Trinity: importance of intrinsic disorder for protein function // Human Genome News. 2002. V. 12. P. 13–14.

Gribskov M., McLachlan A.D., Eisenberg D. Profile analysis: detection of distantly related proteins // Proc. Natl Acad. Sci. USA. 1987. V. 84. P. 4355–4358.

Henikoff S., Henikoff J.G. Performance evaluation of amino acid substitution matrices // Proteins. 1993. V. 17. P. 49–61.

Li X., Obradovic Z., Brown C.J., Garner E.C., Dunker A.K. Comparing predictors of disordered protein genome informatics, 2000. V. 11. P. 172–184.

Salamov A.A., Solovyev V.V. Protein secondary structure prediction using local alignments // JMB. 1997. V. 268. P. 31–36.

Shindyalov I.N., Bourne P.E. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path // Protein Engng. 1998. V. 9. P. 739–747.

MOLECULAR MODELING AND COMPARATIVE ANALYSIS OF AMINO-TERMINAL DOMAIN OF NMDA IONOTROPIC GLUTAMATE RECEPTORS

Belenikin M.S.

Lomonosov Moscow State University, Department of Chemistry, Lenin Hills, 119992, Moscow, Russia, e-mail: bm@org.chem.msu.su

Keywords: *ionotropic glutamate receptor, NMDA, amino-terminal domain, polyamine binding site*

Summary

Motivation: One of the important domains of ionotropic glutamate receptor is amino-terminal domain. However the spatial structure of this domain is unknown. Thus it is necessary to model spatial structure of amino-terminal domain for NMDA ionotropic glutamate receptors.

Results: Using threading and homology methodologies we have constructed the spatial structure of NMDA amino-terminal domain.

Introduction

Glutamatergic system plays a central role in the functioning of central nervous system of mammalian. This system consists of metabotropic and ionotropic glutamate receptors. Metabotropic glutamate receptors (mGlu1–mGlu8) are G-protein coupled receptor, the structure of mGlu1 amino-terminal domain was determinate by X-ray (Kunishima *et al*., 2000; Tsuchiya *et al*., 2002). Ionotropic glutamate receptors (iGlu) are consists of three classes: NMDA (N-methyl-D-aspartate), AMPA (*RS*-2-amino-3-(3-hydroxy-5-methyl-4-isoxazolyl)-propionate) and kainate receptors. For ionotropic glutamate receptors experimentally (by X-ray) have been detected only the structures of glutamate-binding domain of Glu1 (AMPA receptor) (Armstrong, Gouaux, 2000) and identical glycine-binding site of NR1 subunit of NMDA receptor (Furukawa, Gouaux, 2003). Amino-terminal domain is one of the structural domains of iGlu receptors. Antagonists of modulator polyamine binding site has lower activation threshold in compare with competitive antagonists, therefore polyamine binding site structure may help to the design of ligands attenuating over-activation of ionotropic glutamate receptors. However, no experimental data concerning the spatial structure of amino-terminal domain is currently available. The goals of this work are fold identification, molecular modeling of amino-terminal domains and comparative analysis of polyamine binding sites of all NMDA-receptor subtypes.

Methods and Algorithms

At first stage of the study, for the fold identification of amino-terminal domains we applied the method of threading, using the Threader 3.4 software (Jones, Taylor, 1992) and built-in database, to search for potential template proteins for the amino-terminal domains of all iGluR subtypes, both taking and not taking into account the secondary structure predicted by the PsiPred 2.3 method (Jones, 1992). At the second stage of the study, on the based of aminoacid sequence alignments we performed the generations of the amino-terminal domain models using the Modeller 6 (Sali, Blundell, 1990) and Sybyl 6.9 (Clark *et al*., 1989) software. Estimation of the folding quality of the constructed models was made by 3D–1D profiles construction that had positive values (Luthy *et al*., 1992). The local geometry of the protein models was estimated as good using the Procheck software (Laskowski *et al*., 1993).

Results and Discussion

For the fold identification of amino-terminal domains the threading methodology of comparative modeling was used (similarly to our modeling of cysteine-rich domains of mGlu receptors (Belenikin *et al*., 2004)). According to the methodology applied in given work, the calculations are performed for amino-terminal domains of all sixteen (rather than one) iGlu (NMDA: NR1, NR2A-D, NR3A-B; AMPA: Glu1-4 and Kainate: Glu5-7, KA1-2) receptor subunits. The analysis of the data revealed potential leader folding patterns with high accuracy – pariplasmic binding protein-like, class – alpha and beta proteins (a/b). Used complex approach and calculations for all receptor subtypes allowed us to decrease the probability of errors during computed identification of protein folding pattern and to increase accuracy of modeling of ligand binding centers. Structure of ligand-binding core is in compliance with a number of experimental data on site-directed mutagenesis.

The tertiary structure of amino-terminal domains is characterized by the presence of two complementary lobes (each consisting of β-sheets surrounded by α-helices), which are linked by a labile articulated fragment. It is similar to the tertiary structure of the functionally related glutamate-binding proteins.

Thus on the final stage comparative analysis of structure of polyamine binding sites for different subtypes of ionotropic glutamate and studying of protein ligand interactions for a number of noncompetitive ligands were made.

References

Armstrong N., Gouaux E. Mechanisms for activation and antagonism of an AMPA-sensitive glutamate receptor: crystal structures of the GluR2 ligand binding core // Neuron. 2000. V. 28. P. 165–181.

Belenikin M.S., Palyulin V.A., Zefirov N.S. The modeling of the structure of the cysteine-rich domain of Metabotropic Glutamate Receptor // Doklady Biochem and Biophys. 2004. V. 394. P. 21–25.

Clark M., Cramer R.D., Van Opdenbosch N. Validation of the general purpose tripos 5.2 force field // J. Comp. Chem. 1989. V. 10. P. 982–1012.

Furukawa H., Gouaux E. Mechanisms of activation, inhibition and specifity: crystal structures of the NMDA receptor NR1 ligand-binding core // EMBO J. 2003. V. 22. P. 2873–2885.

Jones D.T., Taylor W.R., Thornton J.M. A new approach to protein fold recognition // Nature. 1992. V. 358. P. 86–89.

Jones D.T. Protein secondary structure prediction based on position-specific scoring matrices // J. Mol. Biol. 1999. V. 292. P. 195–202.

Kunishima N., Shimada Y., Tsuji Y., Sato T., Yamamoto M., Kumasaka T., Nakanishi S., Jingami H., Morikawa K. Structural basis of glutamate recognition by a dimeric metabotropic glutamate receptor // Nature. 2000. V. 407. P. 971–977.

Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M. PROCHECK – a program to check the stereochemical qualite of protein structures // J. Appl. Crystallogr. 1993. V. 26. P. 283–291.

Luthy R., Bowie J.U., Eisenberg D. Assessment of protein models with three-dimensional profiles // Nature. 1992. V. 356. P. 83–85.

Sali A, Blundell TL. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming // J. Mol. Biol. 1990. V. 212. P. 403–428.

Tsuchiya D., Kunishima N., Kamiya N., Jingami H., Morikawa K. Structural views of the ligand-binding cores of a metabotropic glutamate receptor complexed with an antagonist and both glutamate and Gd^{3+} // Proc. Natl Acad. Sci. USA. 2002. V. 99. P. 2660–2665.

KINETICS OF PROTEIN FOLDING: LATTICE SIMULATIONS AND ANALYTIC MODEL

Chekmarev S.F.*[1], Krivov S.V.[2], Karplus M.[2,3]

[1] Institute of Thermophysics SB RAS, 630090 Novosibirsk, Russia; [2] Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 67000 Strasbourg, France; [3] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA
***** Corresponding author: e-mail: chekmare@itp.nsc.ru

Keywords: *lattice heteropolymer, simulations, folding kinetics, analytic model*

Summary

Motivation: Although an essential progress in the understanding of protein folding has been achieved with the "landscape viewpoint", the mean force surfaces do not give a complete description of folding kinetics. One source of complementary information is the folding time distributions, which are directly related to the kinetics and can, in principle, be measured in protein folding experiments. The goal of the present work is to examine the relation between these distributions and kinetics in detail.

Results: A 27-residue lattice heteropolymer subject to Monte Carlo dynamics on a simple cubic lattice is studied for a range of temperatures, with a specific attention given to the folding time distributions. The results are compared with those from methods based on mean force surfaces expressed in terms of a reduced set of variables (the "landscape viewpoint"). To get an insight into folding kinetics, a simple kinetic model is formulated that describes the transitions between characteristic states of the heteropolymer. These states correspond to an unfolded chain, semicompact random globule, dead-end traps, and the native state. The comparison of the theoretical distributions found from this model with the simulation results made possible to determine the rate constants for the essential channels of folding process and analyze the importance of these channels with temperature.

Introduction

A full understanding of the way in which proteins fold into their functional three-dimensional native structures has not been yet achieved although considerable progress has been made recently [1]. Joined efforts of many research groups have resulted in a "new view" of protein folding, based on the landscape viewpoint and funnel concept [2, 3], which has provided an essential progress in the understanding of protein folding. At the same time, the mean force surfaces, depending on a reduced set of the variables, such as the number of total and native contacts, do not give a complete description of the kinetics because none set of a few variables is generally able to describe a folding path of the system through the multidimensional conformation space. One source of complementary information is the folding time distributions, which determine the probability that the system, being in an unfolded state, reaches the native state in a given time. Although it has been repeatedly indicated that the folding time distributions vary considerably depending on the folding scenario (from a single-exponential distribution for the two-state kinetics to multiexponential ones for more complex scenarios), their relation to folding kinetics has not been examined in depth.

Lattice simulations

As the model system, we consider the widely studied 27-residue heteropolymer on a cubic lattice, in which the interaction energy was selected from separate Gaussian distributions corresponding

to native and nonnative contacts and also took into account hydrophobic effects [4]. This model is simple enough to permit many folding trajectories to be calculated, yet realistic enough that it can represent a polypeptide chain, with the number of possible conformations being larger than that, which can be sampled in a folding trajectory. The native (lowest energy and free energy) state of the heteropolymer is a fully compact 3 x 3 x 3-cubic structure with $N = N_{nat} = 28$, where N and N_{nat} is the number of total and native contacts, respectively (Fig. 1).

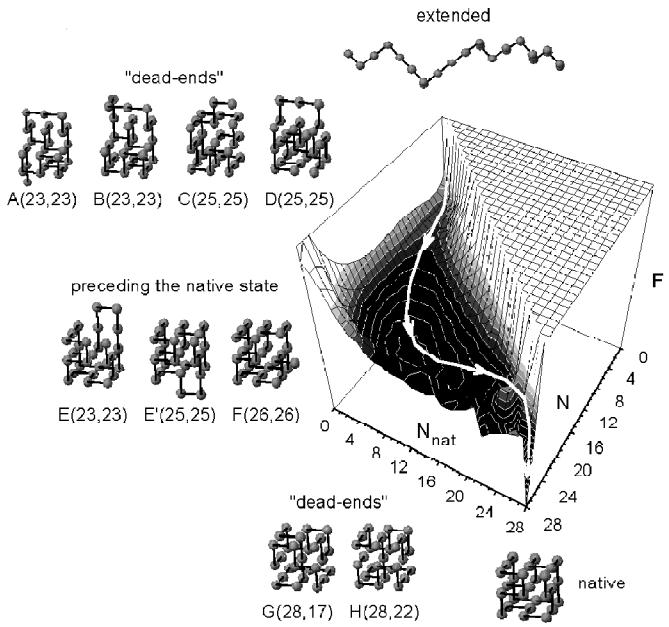


Fig. 1. Schematic free energy surface for folding of the 27-bead heteropolymer. Structures A to H present characteristic configurations of the heteropolymer. In brackets there are indicated the numbers of total and native contacts.

To simulate folding process, we used the Monte Carlo (MC) dynamics. Three types of moves were allowed; they are end flips, corner flips, and two-bead crankshaft rotations. The trajectories were started from a randomly generated unfolded chain and terminated upon reaching the native state, so that the folding time represents a first passage time (Fig. 1). Collecting the simulation data for the ensemble of trajectories, we calculated the probability for the system to be in a state with N and N_{nat} contacts, and also the mean energy of the system at this point. These data were used to build the residence probability, mean potential energy, free energy and entropy surfaces as functions of N and N_{nat}.

The analysis of the folding trajectories mapped the mean force surfaces has shown that the heteropolymer folds roughly in two stages, in accordance with the previous studies. In the first stage, a semicompact random globule is formed, to which a broad basin with a large total number of contacts and a moderate number of native contacts corresponds in the free energy surface. In the second stage of the folding process, this semicompact globule rearranges and makes a conformational search for the native structure ($N=N_{nat}=28$). Also, it has been found that some minima on the free energy surface represent “dead-end” (DE) traps, from which the system can not proceed to the native state without a return to the semicompact globule basin. Some configurations representing the most essential minima are shown in Fig. 1.

Kinetic Model

To analyze the folding kinetics, we used a simple model describing the transitions between the characteristic states of the heteropolymer

$$\frac{dn_{\mathrm{u}}}{dt} = -r_{\mathrm{gu}} n_{\mathrm{u}} \tag{1}$$

$$\frac{dn_{\mathrm{g}}}{dt} = r_{\mathrm{gu}} n_{\mathrm{u}} + r_{\mathrm{gd}} n_{\mathrm{d}} - (r_{\mathrm{fg}} + r_{\mathrm{dg}}) n_{\mathrm{g}} \tag{2}$$

$$\frac{dn_{\mathrm{d}}}{dt} = r_{\mathrm{dg}} n_{\mathrm{g}} - r_{\mathrm{gd}} n_{\mathrm{d}} \tag{3}$$

$$\frac{dn_{\mathrm{f}}}{dt} = r_{\mathrm{fg}} n_{\mathrm{g}} \tag{4}$$

Here the subscripts u, g, d and f label the unfolded chain, globule, DE, and the native state, respectively, n_α is the population of a state, and $r_{\beta\alpha}$ is the rate constant for the transitions from α to β state. The folding time distribution is determined as $p_{\mathrm{f}}(t) = dn_{\mathrm{f}} / dt = r_{\mathrm{fg}} n_{\mathrm{g}}$.

Results and Discussion

The analytical solution obtained from the kinetic model showed that the steep rise at short times corresponds to the transitions from an unfolded chain to the random globule, and the decay at longer times to those from the globule to the native state. In particular, if the decay curve is double-exponential, which is observed at $T < T_{\mathrm{f}}$ (Fig. 2a), where T_{f} is the optimal folding temperature, associated with the minimum folding time, the first, rapidly decreasing exponential term corresponds to the folding trajectories that do not visit the DEs, and the following it, slowly decreasing term to the trajectories that do this, possible returning repeatedly to the globule state until reaching the native state.

Among other things, the analytical solution has led to several not obvious conclusions of general character that are well supported by the simulations. We found that if the DE exists but it is in equilibrium with the globule (due to a fast rate of return from the DE and globule), the folding time distribution is a single-exponential at longer times, which is characteristic of the absence of the DE traps. In this case, the globule and DEs constitute an extended globule, and the transition from the globule to the native state is characterized by a single, effective rate constant, which takes into account the rates of the interconversion between the DEs and globule. It thus follows that a single exponential decay of the folding time distribution at intermediate and longer time scales does evidence that the folding process obeys a two-state kinetics, but not necessarily that off-pathway minima are absent. Another striking result is that the mean folding time (MFT) is independent of the height of the barrier between the globule and DE regions, with the contribution of these regions completely determined by the ratio of their partition functions.

One essential property of the folding time distributions is that the rate constants for the principal channels of transitions can be found by comparing these distributions with those from the kinetic model. Using this approach, we were able to determine the rate constants for a wide temperature

range and analyze how the significance of different channels of transition and off-pathway minima varies with the temperature. The analytical solutions with the rate constants determined from the simulations are shown in Figure 2 by solid lines.

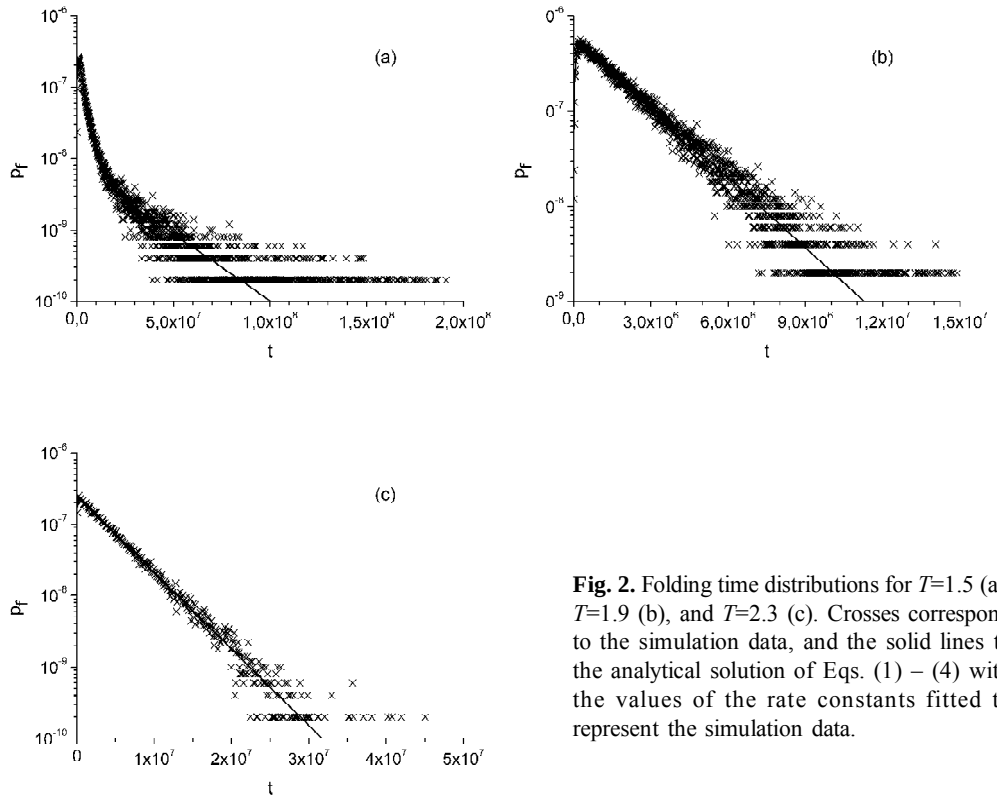


Fig. 2. Folding time distributions for T=1.5 (a), T=1.9 (b), and T=2.3 (c). Crosses correspond to the simulation data, and the solid lines to the analytical solution of Eqs. (1) – (4) with the values of the rate constants fitted to represent the simulation data.

Acknowledgements

This work was supported by a grant from the INTAS (#2001-2126). S. Ch. also acknowledges a support from the RFBR (# 02-03-32048) and SB RAS (#119).

References

1. Dinner A.R., Šali A., Smith L.J., Dobson C.M., Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment // Trends Biochem. Sci. 2000. V. 25. P. 331–339.
2. Onuchic J.N., Luthey-Schulten Z., Wolynes P.G. Theory of protein folding: the energy landscape perspective // Annu. Rev. Phys. Chem. 1997. V. 48. P. 545–600.
3. Šali A., Dobson C.M., Karplus M. Protein folding: A perspective from theory and experiment // Angew. Chem. Int. Ed. 1998. V. 37. P. 869–893.
4. Šali A., Shaknovich E., Karplus M. How does a protein fold? // Nature. 1994. V. 369. P. 248–251.

IDENTIFICATION AND ANALYSIS OF CELL SURFACE NUCLEIC ACIDS-BINDING PROTEINS

Chelobanov B.P.*, Ivanisenko V.A.[1], Kharkova M.V., Laktionov P.P., Rykova E.Yu., Vlassov V.V.

Institute of Chemical Biology and Fundamental Medicine SB RAS, Russia; [1] Institute of Cytology and Genetics SB RAS, Russia

***** Corresponding author: e-mail: chelobanov@niboch.nsc.ru

Keywords: *nucleic acids receptor, isolating of nucleic acids binding proteins, DNA- RNA-binding site prediction, protein tertiary structure*

Resume

Motivation: Binding and penetration of nucleic acids into cells was shown to depend on their interaction with the cell surface proteins.

Results: Cell surface NA-binding proteins were identified and analysed.

Introduction

It has been shown that nucleic acids (NA) penetrate into eukaryotic cells by a receptor-mediated process. Consequently, binding of NA to the cell surface proteins is the first step in the determination of the biological activities of nucleic acids, such as immunostimulation and inhibition of gene expression. The cell surface NA binding proteins have been intensively investigated. It has been apparent that large scale proteomic research is required for revealing the mechanisms by which NA penetrates into cells and implements its biological potentialities. A proteomic approach requires the development of novel methodology for the isolation of the proteins involved in NA recognition. We have previously developed a method for isolation of ODN-binding proteins based on affinity modification of cell surface nucleic acid binding proteins with hapten-modified reactive ODN derivatives followed by isolation of the ODN-protein complexes by anti-hapten affinity chromatography (Laktionov *et al*., 2003).

Current investigation concerns the development of a novel approach for the isolation of the cell surface nucleic acid-protein complexes. The nucleoprotein complexes were isolated taking advantage of the capability of glass surface to bind NA. Isolated proteins were sequenced and identified with database search. Bioinformatics approaches were used for analysis of the nucleic acids binding properties of the isolated proteins.

Materials and Methods

Cells of human epidermoid carcinoma cell line A431 were grown in DMEM medium with 10 % heat inactivated fetal bovine serum at 37 ^{o}C, 5 % CO_2. To remove the NA-protein complexes from the cell surface, cells were washed for 3 minutes at room temperature with phosphate buffered saline containing 5 mM EDTA (PBS/EDTA). The NA-protein complexes were absorbed on glass-milk in 10 mM Tris-HCl, pH 5.5, containing 0.5 % Tween 20. Bound proteins were eluted from glass milk with extensive washing of glass-milk with 10 mM Tris-HCl, pH 5.5, containing 0.5 % Tween 20. The NA-protein complexes were eluted with DNA elution buffer (Laktionov *et al*., 2002) and separated by SDS-PAGE. The range of the protein molecular masses in the PAAG fragments was determined. Each fragment was treated with trypsin separately as described (Shevchenko *et al*., 1996), the peptides obtained were sequenced by MALDI-TOF and analyzed with the MS-Fit program.

The DBS-PRED program was used to predict DNA-binding sites in proteins from their sequence information and using information from known binding sites of proteins (Ahmad *et al*., 2004).

PDBSiteScan provides automated search of the three-dimensional (3D) protein fragments similar in structure to known DNA- and RNA-binding sites (Ivanisenko *et al*., 2004).

Results and Discussion

Binding of nucleic acids with the surface of glass milk and glass filters in the presence of chaotropic salts is widely used for isolation of NA (Chomczynski, 1993). We demonstrated that chaotropic salts are not required for DNA to bind to GM: not less than 25 % of DNA was found to be bound to GM in 10 mM Tris-HCl, pH 5.5 with 0.5 % Tween 20. It was shown that GM binds NA binding protein laktoferrin in the presence of DNA in the same buffer, not in the absence of DNA (Fig. 1).

We have previously demonstrated that cells grown in culture release DNA and RNA that accumulate in culture medium and at the cell surface (Morozkin *et al*., 2004).

The NA-protein complexes were isolated from PBS-EDTA eluate of A431 cells (Fig. 2, lane 1) with a GM protocol. Isolated proteins differ from those of the initial PBS-EDTA eluate (Fig. 2, lane 2).

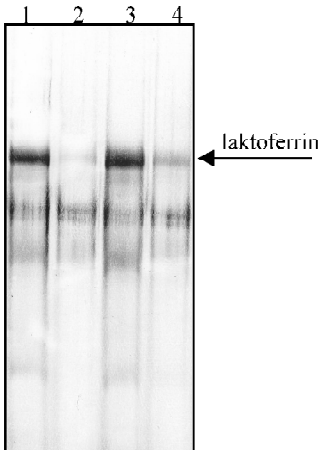


Fig. 1. NA-proteins complexes were absorbed on glass milk in 10 mM Tris-HCl, pH 5.5, containing 0.5 % Tween 20 (Lanes 1, 2) or without Tween 20 (Lanes 3, 4).
Lanes 1, 3 – laktoferrin binding in the presence of DNA;
Lanes 2, 4 - laktoferrin binding in the absence of DNA;
The NA-protein complexes were eluted with DNA elution buffer, separated by SDS-PAGE, transferred onto nitrocellulose and stained with colloidal silver.

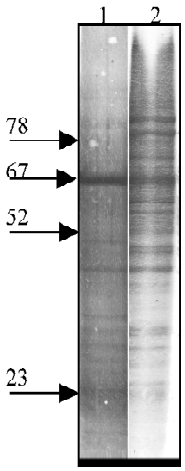


Fig. 2. Isolation of NA-protein complexes eluted from the surface of A431 cells.
Lane 1. – proteins eluted from cell surface with PBS/EDTA. Lane 2. – proteins bound the glass-milk The proteins were separated with 10–20% SDS-PAGE, transferred onto nitrocellulose membrane and stained with colloidal silver.

Isolated NA binding proteins were sequenced and identified with database search. Heat shock protein HSP 90-beta, Heat shock protein HSP 90-alpha, Moesin, Bovin Serum albumin precursor, Cytokeratin 1, Flotillin-1, Phosphoglycerate kinase 1, Ethanolamine kinase-like protein EKI2, Annexin A2, Glyceraldehyde 3-phosphate dehydrogenase, Tropomyosin beta chain, 28 kDa Golgi SNARE protein and probable RNA-dependent helicase p72 were identified with high MOWSE Score reported by the MS-Fit program, which is based on the scoring system in (Pappin *et al*., 1993). Some of these proteins have been previously isolated and identified as NA-binding proteins. Phosphoglycerate kinase 1 and Glyceraldehyde 3-phosphate dehydrogenase were isolated from yeast membrane extract on tRNA-Sepharose (Liu *et al*., 2000). Glyceraldehyde 3-phosphate dehydrogenase was also isolated from nuclear extract of HeLa cells by affinity chromatography on $p(N)_{16}$-Ultrogel A2 (Griffoni *et al*., 2001), Moesin and Ezrin were identified using Southern-Western blotting and identified by mass

spectrometry after separation in 2-D gel (Tschakarjan *et al*., 1999). Cytokeratin 1 was isolated using an affinity modification of intact A431 cells with a oligonucleotide derivative bearing fluorescein on the 5'-end and oxidized uridine on the 3'-end with subsequent isolation modified proteins by affinity chromatography on a column containing anti-Flu-antibodies immobilized on resin (Chelobanov *et al*., 2003); bovine serum albumin modified with ODN derivatives was isolated from cell lysate using anti-BSA IgG after modification of the cell surface proteins with ODN derivatives (Geselowitz, Neckers, 1995). The agreement between the isolated proteins and known DNA-binding proteins support the capability of the GM approach for isolation of DNA-binding proteins for proteomic research. As a result of the combined analysis of the primary and tertiary structures of the identified proteins, using the DBS-PRED and PDBSiteScan programs, we revealed the potential NA-binding sites.

The 3D structures of the potential complexes of DNA and RNA with serum albumin (Fig. 3), phosphoglycerate kinase 1 and glyceraldehyde-3-phosphate dehydrogenase (data not shown) were obtained using the PDBSiteScan program. As Fig. 3 shows serum albumin interacts with DNA through contacts of the alpha-helix with major DNA groove. The data support the NA-binding capabilities of the proteins and will be further used for a detailed simulation of the protein-nucleic acid complexes using molecular dynamics.

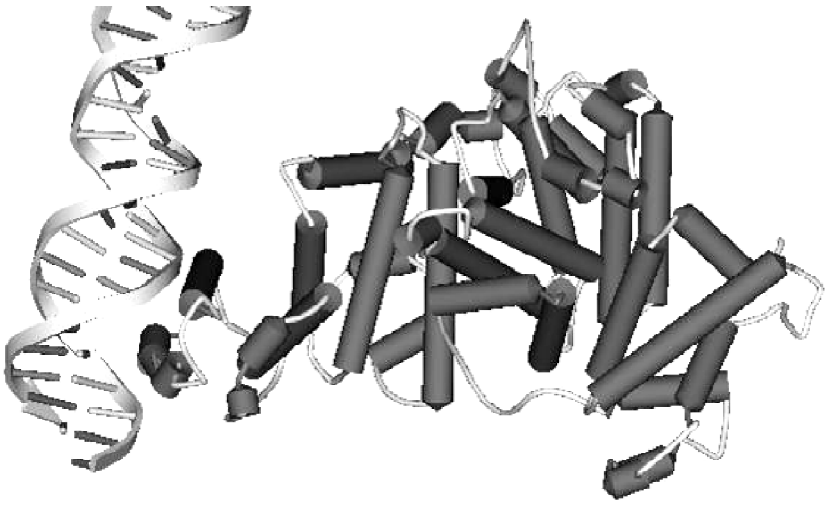


Fig. 3. Potential complex of serum albumin with DNA. In building the complex, atom coordinates from the PDB databank (ID 1E7E) for protein and (ID 1H88) for DNA were used. Protein is depicted as a cylindrical model, DNA as arrows-ladder.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (RFBR, project No. 03-04-48647-a, 01-07-90376 and 03-07-96833-p2003), grant support from Science – Technical Programs of RF Ministry of Education UR.07.01.008, Scientific schools grant SS-1384.2003.4, in part by Award No. NO-008-X1 of CRDF, the Basic Research and Higher Education (BRHE) program NO-008-X1 and grant from Phisico-Chemical Biology program of RAS; the Siberian Branch of the Russian Academy of Sciences (Integration Project No. 119); Russian Ministry of Industry, Science and Technologies (grants No. 43.073.1.1.1501).

References

Ahmad S., Gromiha M.M., Sarai A. Analysis and Prediction of DNA-binding proteins and their binding residues based on composition, sequence and structure information // Bioinformatics. 2004. V. 20. P. 477–486

Chelobanov B.P., Laktionov P.P., Kharkova M.V., Rykova E.Y., Pyshnyi D.V., Pyshnaya I.A., Marcus K., Meyer H.E., Vlassov V.V. Interaction of keratin k1 with nucleic acids on the cell surface // Biochemistry. (Mosc.). 2003. V. 68. P. 1239–1246.

Chomczynski P. A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples // Biotechniques. 1993. V. 15. P. 532–537.

Geselowitz D.A., Neckers L.M. Bovine serum albumin is a major oligonucleotide-binding protein found on the surface of cultured cells // Antisense Res. Dev. 1995. V. 5. P. 213–217.

Griffoni C., Laktionov P.P., Rykova E.Y., Spisni E., Riccio M., Santi S., Bryksin A., Volodko N., Kraft R., Vlassov V., Tomasi V. The Rossmann fold of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) is a nuclear docking site for antisense oligonucleotides containing a TAAAT motif // Biochim. Biophys. Acta: Mol. Cell. Biol. Lipids. 2001. V. 1530. P. 32–46.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins // Nucleic Acids Res. 2004. 32, in press.

Laktionov P.P., Chelobanov B.P., Kharkova M.V., Rykova E.Yu., Pyshnyi D.V., Pyshnaya I.A., Marcus K., Meyer H.E., Vlassov V.V. Ñell surface oligonucleotide-binding proteins of human squamous carcinoma À431cells // Nucleosides Nucleotides Nucleic Acids. 2003. V. 22. P. 1715–1719.

Laktionov P., Tamkovich S., Simonov P., Rykova E., Vlassov V. Method of isolation of deoxyribonucleic acids // Russian patent 2002126328, October, 2. 2002.

Liu Q.S., Jin Y.X., Jiang D.S., Liu J.H., Wang T.P. Isolation and identification of two novel transfer RNA binding proteins from yeast membrane // 18th tRNA workshop “tRNA 2000”. Cambridge, UK. 2000.

Morozkin E.S., Laktionov P.P., Rykova E.Y., Vlassov V.V. Release of DNA and RNA by eukaryotic cells in culture. Ann N Y Acad Sci. 2004. (in press).

Pappin D.J.C., Hojrup P., Bleasby A.J. // Curr. Biol. 1993. V. 3. P. 327–332.

Shevchenko A., Wilm M., Vorm O., Mann M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels // Anal. Chem. 1996. V. 68. P. 850–58.

Tschakarjan E., Trappmann K., Immler D., Meyer H., Mirmohammadsadegh A., Hengge U.R. Keratinocytes take-up naked plasmid DNA: Evidence for DNA binding proteins in keratinocyte membranes // International conference, Israel, Jerusalem. 1999.

EXPERIMENTAL AND COMPUTER EVALUATION OF THE ABILITY ssT-DNA-BINDING VirE2 PROTEIN TO INTERACT WITH LIPID MEMBRANES

Chumakov M.I.*, Burmatov A.V., Bogdanov V.I., Volokhina I.V.

Institute of Biochemistry and Physiology of Plants and Microorganisms, RAS, 13 Pr. Entuziastov, Saratov 410049, Russia

***** Corresponding author: e-mail: chumakov@ibppm.sgu.ru

Keywords: *VirE2 protein, ssT-DNA, lipid membranes, transfer, porins, computer simulation*

Summary

Motivation. The aim of this work was to study VirE2-protein-artificial membrane interactions and to perform a computer simulation of VirE2-protein pore-forming capacity.

Results. The interaction of a VirE2-protein-ssT-DNA complex with flat black membranes was investigated. Using the "patch-clamp" method, we revealed for the first time of a single, long-time (up to 7 s) jumps of membrane conductivity (or channel formation) during coincubation with VirE2 protein in a voltage-dependent manner. It is possible that VirE2 protein forms a membrane pore for T-DNA transfer. We used porins (as β-sheet containing) and bacteriorodopsin (as α-sheet containing) transmembrane proteins as models for computer comparison with VirE2 protein. Using an original program, "Graphic Imagination", we predicted from 3 to 6 β-sheets for aquaporins and up to 15 β-sheets for VirE2 protein and found most similarity of VirE2 protein at this trait with nucleoporins. But we did not find typical transmembrane segments for VirE2 protein by using the MEMSAT program.

Availability : The software application for Windows is available upon request from the authors.

Introduction

Members of the genus *Agrobacterium* can produce undifferentiated tumors on a wide range of dicotyledonous plants. It was shown that *A. tumefaciens* separately transfers VirE2/VirE1 and ssT-DNA-VirD2 complexes to the plant cell cytoplasm where the so called "T-complex" is formed (Gelvin, 1998).

It is assumed that VirE2 proteins cover T-DNA in a single-stranded (ss) DNA-binding protein manner, form a membrane-spanning pore or a channel, and increase the membrane conductivity and promote short peptides translocation across the membrane (Dumas *et al*., 2001). The aim of this work was to study VirE2-protein-artificial membrane interactions and to perform computer simulation of VirE2-protein pore-forming capacity.

Methods and Algorithms

Protein VirE2 was isolated from cells of *E. coli* strain XL1-blue, containing recombinant plasmid pQE31-*virE2*. Purification of a 6-(His)-target recombinant protein was made from sonicated bacterial cells by affinity chromatography on an Ni-NTA-agarose column. The isolated VirE2 protein was tested by immunodot using polyclonal antibodies obtained to VirE2-protein.

MPsrch, GenTHREADER and mGenTHREADER programs located at (http://bioinf.cs.ucl.ac.uk/psipred/) were used to searching for homologous proteins to agrobacterial VirE2 protein in the SWISS-PROT database. The MEMSAT program we used to search for transmembrane segments in VirE2 protein.

We prepared original programs "Conversion" and "Graphic Imagination", for visualization and adaptation of data from the PSIPRED server. The program "Conversion" is intended for the transformation of the results of prediction of protein secondary structure that were obtained from the PSIPRED server into the Graph program format. This format is a structured text file (of the *.txt type).

The first line is of a descriptive character. It may contain any symbols and is intended for the description of the structure of the protein in the file. The second and subsequent lines contain the prediction itself. The line's structure: first symbol, prediction probability (0-9); second symbol, the predicted structure (C, curve; H, α helical; E, β sheet); third symbol, amino acid (one-letter code).

Implementation and Results

It seems that the T-DNA could be transferred into eukaryotic cells in the following way: (a) during close contact (lipid membrane fusion or through VirE2-depended pores in plant membranes); (b) by excretion from the bacterial cell and following adsorption and endocytosis by the plant cell; (c) "injection" through special extracellular transfer structures after docking the bacterial cell to the plant cell surface.

It is belived that increasing the membrane conductivity can promote short peptides (and possible T-DNA) translocation across the membrane (Dumas *et al*., 2001). We investigated the interaction of VirE2-protein with flat black membranes by "patch-clamp" methods. After formation of a lipid bilayer, the purified VirE2-His6 and ssT-DNA were added to the chamber, and an electrical potential difference of 10–300 mV was applied across the membrane. Using the "patch-clamp" method, we observed an increase in the bilayer conductivity during co-incubation with VirE2-protein in a voltage-dependent manner. We revealed for the first time single, long-time (up to 7 s) jumps (or channels) of membrane conductivity (Fig. 1).

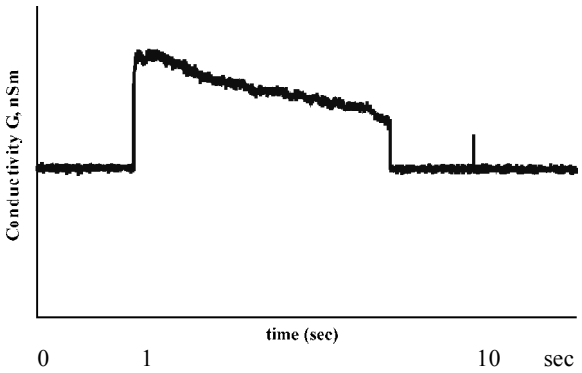


Fig. 1. Change in artificial lipid membrane conductivity during incubation with VirE2 protein (1 min after injection of 16 ng/µl VirE2).

We did not find any proteins homologous to agrobacterial VirE2 protein in the SWISS-PROT database. Since the tertiary structure for this protein unknown, we used known transmembrane proteins like porins (as ß-sheet containing) and bacterio-rodopsin (as α-sheet containing) as models for computer comparison with VirE2 protein. Using the "Graphic Imagination" program we predicted from 3 to 6 ß-sheets for aquaporins and up to 15 ß-sheets for VirE2 protein and found most similarity of VirE2 protein with nucleoporins at this trait (Fig. 2). But we did not find typical transmembrane segments for VirE2 protein using the MEMSAT program. On the other hand, this program did not predict any transmembrane segment for nucleoproteins (P49687, P57740, O14310).

Thus, we found that VirE2 protein interacted with the artificial lipid membrane increasing membrane conductivity and found up to 15 β-sheets for VirE2 protein using computer analysis, but we did not observe a typical transmembrane segments for VirE2 by using MEMSAT program.

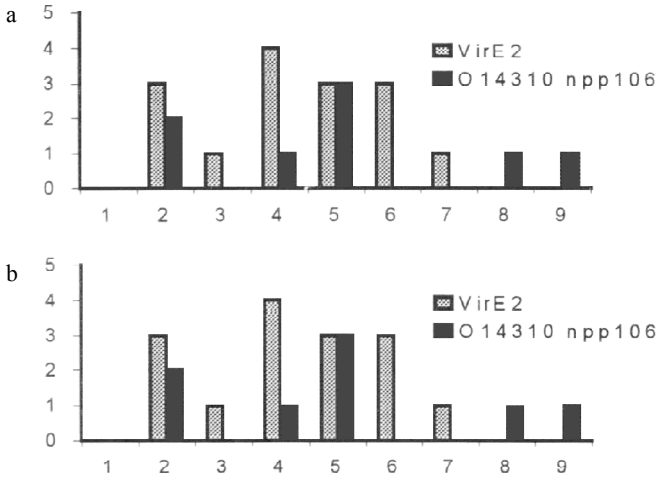


Fig. 2. Distribution of second structures (β-sheet) for VirE2 protein and nucleoporins (P49687(a), 014310(b)) by using "Graphic Imagination" program.

Acknowledgements

We thank G. Maksaev for help with membrane conductivity experiments, Dr.C. Baron for VirE2 antibodies and Yu. Chizmadzev for support.
This research was supported in parts by the Russian Foundation for Basic Research (grants 02-04-49496 and 03-04-48382) and by the Russian Ministry of Education (grant E02-6.0-181).

References

Dumas F., Duckely M., Pelczar P., Van Gelder Patrick, Hohn B. An *Agrobacterium* VirE2 channel for transferred-DNA transport into plant cells // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 485–490.

Gelvin S.B. *Agrobacterium* VirE2 protein can form a complex with T strand in the plant cytoplasm // J. Bacteriol. 1998. V. 181. P. 4300–4302.

MEMBRANE PROTEINS: THE NEW INSIGHTS *via* COMPUTATIONAL EXPERIMENTS

Efremov R.G.*, Volynsky P.E., Nolde D.E., Vereshaga Y.A., Konshina A.G., Simakov N.A., Arseniev A.S.

Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, ul. Miklukho-Maklaya 16/10, GSP Moscow, 117997 Russia
***** Corresponding author: e-mail: efremov@nmr.ru

Keywords: *theoretical membrane models, Monte Carlo simulation, protein-lipid interactions, molecular dynamics, hydrated lipid bilayers, implicit membrane.*

Summary

Motivation: Membrane and membrane-active peptides and proteins play crucial role in numerous cell processes, like membrane fusion, signaling, ion conductance, and so forth. Many of them act as highly specific and efficient drugs or drug targets, and therefore, attract growing interest in medicine and biotechnology. Because of experimental difficulties with their structural characterization, essential attention is given now to molecular modeling approaches. The main difficulty inherent in these methods is related to correct treatment of protein-lipid interactions. To solve the problem, models with implicit and explicit consideration of membrane-mimic media may be employed.

Results: Here we review our recent applications of both classes of theoretical models in molecular dynamics (MD) and Monte Carlo (MC) simulations of a wide class of peptides and proteins with membranes. Among the studied objects there are fusion, signal, antimicrobial, and other peptides, proteins with different folds and mode of membrane binding (Efremov *et al*., 1999a, b; 2001, 2002a, b). Theoretical background of the membrane models is considered with examples of their applications to biologically relevant problems. Testing against experimental data shows that the calculations give good predictions both for the association state and peptides' (or proteins') orientation relative to the membrane. Perspectives of the modeling techniques are discussed.

Introduction

Key role of membrane proteins (MP) in a cell is determined by their ability to detect and transmit biological signals across lipid bilayers, to realize transmembrane (TM) ion and molecular transport, to facilitate membrane fusion, and so forth. Furthermore, peptides and proteins which may interact with membranes to accomplish their functional activities constitute ~30 % of all proteins encoded by the whole genomes. Therefore, delineation of the structure-function relationships for MPs represents an intriguing challenge in the field of structural biology. Apart from fundamental importance (studies of general principles of protein insertion, folding and stabilization in bilayer), solving the problem is invaluable in the optimization of these molecules' behavior for pharmaceutical and biotechnological applications, such as the development and targeted delivery of drugs through membranes or action on membrane-bound receptors, the design of MPs with prescribed properties, gene therapy, and disease control. Enormous experimental difficulties related to structural studies of MPs call for elaboration of independent techniques destined to solve the problem. Molecular modeling represents a promising alternative, which can considerably extend and complement the possibilities of traditional structural biology tools, like X-ray and NMR spectroscopy

Methods

The spatial structure and the mode of membrane binding for peptides and proteins were assessed *via* unrestrained MC conformational search in the dihedral angles space and/or *via* long-term MD simulations. In the first the case the membrane was represented as a hydrophobic slab described by a solvation potential of mean force (Efremov *et al*., 1999a; 2001). In the second case, several full-atom models of hydrated lipid bilayers and micelles were employed.

Results

A number of new approaches in molecular modeling of complex multi-component membrane and micellar systems and their complexes with proteins and peptides were developed. These techniques were applied to solve a wide class of problems destined to structural, dynamical, and functional characterization of membrane peptides and proteins with different folds (alpha-helical, beta-sheet, alpha/beta-structural) and different mode of membrane binding (transmembranal and peripheral). To take into account heterogeneous membrane environment, two types of membrane models were elaborated. In the first one the influence of lipid bilayer is described in terms of a special energy term added to the potential energy function of a protein. Such methods represent a reasonable compromise between the structureless uniform dielectric media and the full-atom membrane models. They are less computationally expensive than explicit solvent calculations and, therefore, are able to address questions about the structure and function of MPs on rather larger time scales. In models of the second type both, the lipid molecules and the water on either side of the membrane are represented atomistically. Computer simulations of peptides and proteins in full-atom lipid bilayers and detergent micelles provide a wealth of very useful structural and dynamical data about equilibrium properties of these systems, protein-lipid interactions, parameters of binding, protein-induced changes in membrane, and others. On the other hand, the payment for such a detailed view is the large size of simulated objects and, consequently – enormous computational resources required.

The implicit membrane model was applied to study a large number of peptides and proteins: antimicrobial and fusion peptides, signal peptides of the membrane protein PhoE, cardiotoxins from snake venom, TM helices of integral membrane proteins (bacteriorhodopsin, nicotinic acetylcholine receptor, etc.), glycophorin A, and others. The results obtained were shown to be in a good agreement with the experimental data. The membrane model permits correct assessment of a number of phenomena which accompany binding of proteins to membranes and drive protein-protein interactions in lipid bilayers. This extended considerably the class of studied problems, and provides an important insight into the biological functions of the objects under investigation.

Full-atom models of detergent micelles and explicit hydrated lipid bilayers of different content and physico-chemical characteristics (temperatures of the phase transitions, containing zwitterionic and charged, saturated and unsaturated lipids) were elaborated. The systems were studied *via* long-term MD simulations. Optimal computational protocols for MD were developed and tested. They provided good correlation between the simulation results and the experimental data. The membrane models were further employed in MD studies of interactions of a number of peptides and proteins (fusion peptides, cardiotoxins) with different lipid bilayers and detergent micelles. A general conclusion was reached that nowadays the theoretical methods of simulations of peptides and proteins in explicit lipid bilayers – with surrounding water molecules and ions – represent the most realistic approximation for studies of protein-membrane interactions. This class of models conforms reasonably well to available physicochemical data on lipid bilayer structure and processes accompanying binding of small compounds, peptides and proteins. MD and MC techniques give an insight into molecular events on the time scale ~10–100 ns, like the spatial structure, the stability and thermal fluctuations of the polypeptide and the membrane's constituents, the energetic aspects

of their interactions, the ion transport across the bilayer and some others. On the contrary, a number of important "slow" processes, specific for protein-membrane systems, are still difficult or impossible to assess on the current time scales. Among them: the large structural deformations of both the protein and the bilayer, induced by protein insertion (this is especially interesting for studies of membrane destabilization and/or fusion, ion channel formation, and so forth), the cooperative movements related to phase transitions, the dynamics of domains in lipid membranes, the "flip-flop" events, and some others.

One of the most promising way to employ computer simulations in studies of membrane peptides and proteins is a combined usage of implicit and explicit membrane models. In this case low-energy states of a protein found in implicit membrane provide good starting points for subsequent long-term simulations in full lipid bilayers. Shortcomings and perspectives of different strategies in setting up and performing computer simulations of peptides and proteins in membrane-mimic media are discussed.

Acknowledgements

This work was supported in part by the Programme RAS MCB, by the Ministry of Science and Technology of Russian Federation (the State contract No. 43.073.1.1.1508, grant SS-1522.2003.4), and by the Russian Foundation for Basic Research (grant 04-04-48875a). R.G.E. and D.E.N. are grateful to the Science Support Foundation (Russia) for the grants awarded.

References

Efremov R.G., Nolde D.E., Vergoten G., Arseniev A.S. A solvent model for simulations of peptides in bilayers. I. Membrane-promoting α-helix formation // Biophys. J. 1999a. V. 76. P. 2448–2459.

Efremov R.G., Nolde D.E., Volynsky P.E., Chernyavsky A.A., Dubovsky P.V., Arseniev A.S. Factors important for fusogenic activity of peptides: molecular modeling study of analogs of fusion peptide of influenza virus hemagglutinin // FEBS Lett. 1999b. V. 462. P. 205–210.

Efremov R.G., Volynsky P.E., Nolde D.E., Arseniev A.S. Implicit two-phase solvation model as a tool to assess conformation and energetics of proteins in membrane-mimic media // Theor. Chem. Acc. 2001. V. 106. P. 48–54.

Efremov R.G., Volynsky P.E., Nolde D.E., Dubovskii P.V., Arseniev A.S. Interaction of cardiotoxins with membranes: a molecular modeling study // Biophys. J. 2002a. V. 83. P. 144–153.

Efremov R.G., Volynsky P.E., Nolde D.E., van Dalen A., de Kruijff B., Arseniev A.S. Monte Carlo simulations of voltage-driven translocation of a signal sequence // FEBS Lett. 2002b. V. 526. P. 97–100.

COMMON STRUCTURAL FEATURES OF HOMEODOMAINS AND HOMEODOMAIN-DNA COMPLEXES

Ershova A.S., Alexeevski A.V.*[1], Spirin S.A.[1], Karyagina A.S.

Institute of Agricultural Biotechnology, Russian Academy of Agricultural Sciences, Moscow, Russia;
[1]Belozersky Institute, Moscow State University, Moscow, Russia
***** Corresponding author: e-mail: aba@belozersky.msu.ru

Keywords: *homeodomain, DNA-binding domain, 3D structure, transcription factor, geometric core, hydrophobic core, hydrogen bond, water mediated contact*

Summary

Motivation: The amount of currently available structural information about DNA-protein complexes allows to reveal the principles of specific site recognition by regulatory and other DNA-binding proteins by the comprehensive comparative analysis of structures. Homeodomains represent one of the best structurally characterized family of DNA-binding proteins. We have performed the analysis for the structures of homeodomains and homeodomain-DNA complexes.

Results: A procedure for the detection of common structural features of a well structurally characterized protein family was developed. The procedure is based on standard computer tools as well as home-made software. The procedure was applied to the homeodomain family. The geometric core and the conserved hydrophobic core of homeodomains were determined. The conserved contacts between homeodomain and DNA, including hydrogen bonds, water mediated contacts, and hydrophobic clusters on the interface, were described.

Introduction

Homeodomain is a common name for a family of homologous ~60 amino acid residue DNA-binding domains of eukaryotic transcription factors. Structurally, it consists of 3 α-helices and the N-terminal "arm". A lot of information on homeodomain sequences and structures was published since the discovery of homeobox (homeodomain coding sequence) (McGinnis *et al*., 1984) . The comparative analysis of the homeodomain-DNA recognition on the base of structural information about 7 homeodomain-DNA complexes was performed by Billeter (Billeter, 1996). At present, much more X-ray and NMR data on homeodomains and homeodomain-DNA complexes became available: 40 entries of Protein Data Bank (PDB) contains 62 structures of 21 different homeodomains from 12 subfamilies; among them there are 49 structures of homeodomain-DNA complexes from 27 PDB entries.

For each solved complex, structural features of homeodomain-DNA interaction were described in the original paper(s) presenting the structure deciphering data. Among these features are hydrogen bonds between protein and DNA, hydrophobic and van der Waals contacts, water bridges, etc. To reveal conserved (therefore, putatively more important) features, it is needed to analyze all data by an uniform approach.

We developed an unified procedure for a comprehensive comparative analysis of protein-DNA complexes. The found conserved features in homeodomain family can be used for the prediction of DNA-binding specificity and for protein engineering experiments. Also, we have used the results of the analysis in designing an automatic detector of homeodomain structures in a PDB entry (under construction).

Methods and Algorithms

Both freely available (RasMol, SwissPDBviewer, GeneDoc) and home-made software was used. The latter includes: CluD, the program for detection hydrophobic clusters in 3D structures (http:

//math.belozersky.msu.ru/~mlt/HF_page.html), Life Core, the program for detection a geometric core in a family of 3D structures (http://www.dpidb.belozersky.msu.ru:8080/), and a program for detection hydrogen bonds between nucleic acid and protein (http://www.dpidb.belozersky.msu.ru/form.htm). All programs were used with the default parameters. The wide-spread sequence-based classification of homeodomains into subfamilies was used (Burglin, 1994). Homeodomains with known 3D structure used in the work belong to the following families: Antp (6;5), En (4;3), Eve (2;2), MatA1 (4;3), Matα2 (7;6), Msh (1;1), NK-2 (1;0), PAX/Prd (3;3), POU (9;6), Tale (4;3). Here in parentheses are the number of available structures and, after semicolon, the number of structures with deciphered coordinates of water molecules.

Implementation, results and discussion

All available structures of homeodomains and homeodomain-DNA complexes were analyzed by a specially developed procedure. The procedure includes (1) a unified numbering of amino acid residues, as well as of DNA bases within the recognition site (Ledneva *et al*., 2001); (2) the superimposition of all available structures; (3) computing a geometrical core of the homeodomain family, i.e., a set of similarly disposed in all structures C_δ atoms (Fig. 1A) (see the exact definition in http://www.dpidb.belozersky.msu.ru:8080/CoreProc/help.html); (4) computing the hydrophobic core for each homeodomain and determining the conserved hydrophobic core of the whole family (Fig. 1B); (5) computing hydrophobic clusters on DNA-protein interface for each complex and detecting clusters that are conserved in specific subfamilies; (6) computing protein-DNA hydrogen bonds (HB) in each complex, classifying them into 3 types: HB with the DNA bases in the major groove, HB with bases in the minor groove, and HB with phosphates; detecting those HB that are conserved in the family or in specific subfamilies (an example is illustrated by Fig. 1C); (7) computing and classifying water bridges between homeodomain and DNA; detecting groups of conserved water molecules.

The geometrical core of the homeodomain family consists of C_α-atoms of the most residues of the first and third α-helices (Fig. 1A). It means that the first and third helices are very conservatively disposed in space and form a main structural determinant of the family.

Conserved homeodomain hydrophobic core, i.e., the residues from the positions involved in the hydrophobic core in all structures, is an essential part of the hydrophobic core of each homeodomain and is located in the center of the protein domain (Fig. 1B). Almost all C_α-atoms of the residues forming the hydrophobic core are included in the geometric core of homeodomains. The conserved hydrophobic core can be regarded as a significant structural determinant of the family as well.

The conserved DNA-protein contacts in homeodomain-DNA complexes can be subdivided onto two categories: the contacts that are common for all members of the homeodomain family and the contacts that are common for certain subfamilies. The first category includes the absolutely conserved bifurcated hydrogen bond between conserved in homeodomains Asn51 and conserved in homeodomain recognition site adenine A103 (here the unified numeration is used) (Fig. 1C). The second category includes (i) a number of hydrogen bonded contacts with DNA phosphate groups: HB of Arg53 with oxygen atom of phosphate group 200, which is observed in all subfamilies except of Eve and NK-2; HB of Tyr/Lys25 with phosphate group 200 (all except Eve and Mata1); HB of the residue 6 with phosphate group 103 (all except Mata1, Matα2, NK-2, and Tale); HB of the residue 44 with phosphate group 103 (all except Eve, Msh, and POU); HB of Lys55 with phosphate group 101 (all except Eve, Mata1 and Tale); HB of Lys/Arg57 with phosphate group 201; (ii) water mediated contacts of the residues 44 and Trp48 with phosphate 103; water mediated contacts of the residues 26 and Arg53 with phosphate group 199; (iii) water mediated contact of the residue 50 with the DNA base 201 (Fig. 1D). Also, the hydrophobic cluster including residue Ile/Val47 and T104 is observed in all subfamilies except of Matα2 and Tale. In Matα2 and Tale structures, which have Asn residue in the position 47, a hydrophobic cluster including Arg54 (Matα2) or Ile54 (Tale) and T201 is observed.

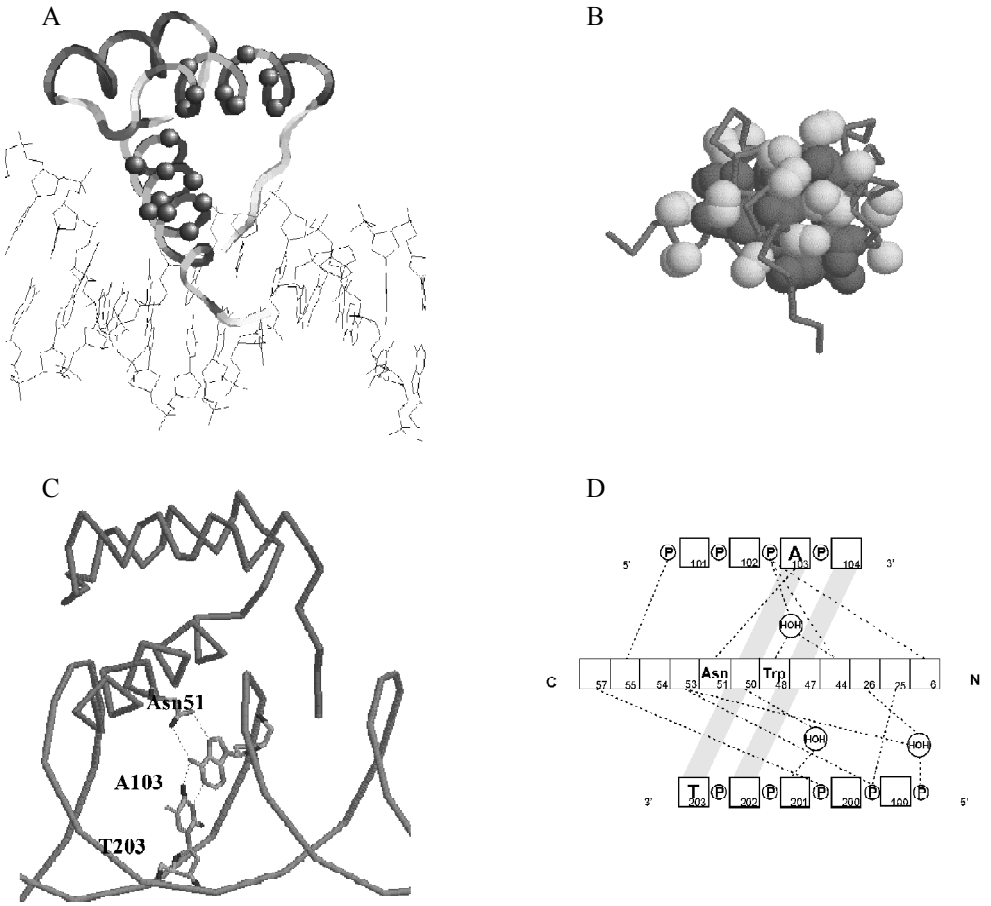


Fig. 1. A – 3D structure of Pit-1 homeodomain in complex with DNA target (PDB code 1AU7). The balls indicate C_α atoms from the geometric core. B – The hydrophobic core of the homeodomain Engrailed (3HDD). The dark-gray balls indicate the carbon atoms involved in the conserved hydrophobic core, the light-gray balls indicate other carbon atoms involved in the hydrophobic core. C – The bifurcated hydrogen bond between Asn51 and A103 in the homeodomain-DNA complex 3HDD. D – A scheme of conserved hydrogen bonds and water-mediated contacts between homeodomain amino acid residues (central row) and DNA direct strand (top row) and reverse strand (bottom row). Each hydrogen bond is realized in at least 7 of 9 homeodomain families. All hydrogen bonds can be realized simultaneously in homeodomain structure (see, for example, 1FJL structure of homeodomain from PAX/Prd family). Dotted lines correspond to hydrogen bonds. Figures A, B, C are made using RasMol 2.7.2.

In addition to three discussed in the text groups of conserved water molecules, we identified nine (!!!) spatially conserved in specific subfamilies interfacial water molecules. These water molecules mediate the contacts with phosphate groups and with DNA bases in the major groove. These data support an important role of water in homeodomain-DNA recognition proposed earlier (see, for example, Billeter, 1996).

We used the obtained data for a functional annotation of key homeodomain residues and formulation of partial 'recognition rules' of homeodomain – DNA interaction (data not shown). The knowledge about homeodomain conserved features can be used also for prediction the DNA recognition sites of homeodomains.

Acknowledgements

The work was partly supported by RFBR (grants 03 04 48476 and 03 07 90157) and Ludwig Institute for Cancer Research (CRDF grant RB0-12771-MO-2).

References

McGinnis W., Levine M.S., Hafen E., Kuroiwa A., Gehring W.J. A conserved DNA sequence in homoeotic genes of the Drosophila antennapedia and bithorax complexes // Nature. 1984. V. 308. P. 428–433.

Billeter M. Homeodomain-type DNA recognition // Prog. Biophys. Molec. Biol. 1996. V. 66. P. 211–225.

Bürglin T.R. A comprehensive classification of homeobox genes // Guidebook to the Homeobox Genes / Ed. D. Duboule. Oxford Univ. Press, 1994. P. 25–72.

Ledneva R.K., Alexeevskii A.V., Vasil'ev S.A., Spirin S.A., Karyagina A.S. Structural aspects of the homeodomain-DNA interaction // Mol. Biol. (Mosk). 2001. V. 35. P. 764–777 (in Russ.).

COMPUTATION OF THE THREE DIMENSIONAL STRUCTURE OF THE HUMAN TYPE (III) COLLAGEN

Filatov I.V.*, Milchevsky Ju.V., Esipova N.G., Tumanyan V.G.

Engelhardt Institute of Molecular Biology RAS, Moscow, Russia
***** Corresponding author: e-mail: ivfilatov@mail.ru

Keywords: *collagen, SNP, three-dimensional structure computation, molecular mechanics*

Resume

Collagen is an important structural protein. The single nucleotide polymorphisms in the collagen genes are responsible for the great number of the diseases [1]. Therefore, it is all-important to understand the collagen three-dimensional structure organization. The objective of the current study is the computation of the three dimensional collagen structures and revelation of the local stability dependence on the sequence.

Introduction

Collagen is the most abundant fibrous protein in mammals. Its makes up to 30 percent of the overall protein weight. Collagen is synthesized by fibroblasts and forms fibers which are responsible for functional integrity of connective tissue. There was found about twenty types of human collagens. Collagen molecule consists of three parallel polypeptide chains that form cross-links. It is well known, that collagen sequence contains glycine residue in the first positions of each tripeptide. Furthermore, collagen is enriched in proline and hydroxyproline. Up till now, however, there have been few structures resolved by X-ray crystallography. The longest of them does not exceed thirty of amino acid residues. This happens because of crystallization difficulties and interpretation complexity of diffraction pattern from collagen structures. Thus, the conformational calculation becomes very important. In order to design a model, it is necessary to define a true type of H-bond net depending on the sequence. The best model for the net of H-bonding, by now, is the so-called Tumanyan-Esipova model [2]. The objective of the current study is modeling of the three dimensional collagen structure based on this model.

Methods and Algorithms

The program ICM was used for conformational calculations [3, 4]. We took a main chain conformation from structures $(GPO)_n$ resolved by X-ray crystallography method [5, 6] as a first approximation for the main chain. The set of rotamers taken from library [7] that provide for the best conformational energy of sampling structures was used as a first approximation. Conformation energy minimization was performed with the help of conformational calculation program ICM.

Results and Discussions

To test our method we carried out calculation of three-dimensional structures that had been resolved by X-ray crystallography earlier. Particularly, we attempted to predict structure of 1BKV protein stored in PDB bank. Resolution for 1BKV was given equal to 2,0 Å. The mean square deviation of results of our computations from experimental data equaled to 1,9 Å. Thus, we demonstrated that the method developed yielded good results and thus calculated the three-dimensional structure of a complete collagen molecule. The calculated structures are in a good agreement with known experimental data. More precisely, the projection of a residue on the helix axis equaled to 2.85 Å for both experimental and computed structures. The helical rotation angle equaled to 36° for the experimental structures and equaled to 38.5° for the computed structures that contain an amino

acid in the second position of tripeptides. Thus we computed structures of octatripeptides describing completely the structure of human collagen III. Also, we computed complete structure of human collagen III with the aid of these structures. We found the most significant parameters that determine the conformational energy of the calculated octatripeptides. The number of hydroxyproline residues in the third position of the triplets and squared mass of the calculated octatripeptade were found as the most significant local stability (see Table).

Table. The most significant parameters that determine local stability

Physical parameters	Regression coefficient	Standard deviation	Fisher`s statistics
Number of the hydroxyproline in the third position	22.8	1.2	362
Squared mass of the calculated octatripeptide	-0.72	0.03	528
Free term 9.5			

Acknowledgements

This work was supported by grants from the Russian Foundation for Basic Research (No. 03-04-49017 and No. 02-04-49114), grant for Ministry of Industry, Science and Technologies (No. 43.071.1.1.1517) and grant on Molecular and Cellular Biology RAS (Program No. 10).

References

1. Milchevsky J.V., Ramensky V.E., Esipova N.G., Tumanyan V.G., Zorov B.S. Molecular modeling of disease-causing single-nucleotide polymorphisms in collagen // SAR QSAR Environ Res. 2001. V. 12(4). P. 383–99.
2. Tumanyan V.G., Esipova N.G. The structure of collagen with a new methodology for modeling two nets of interpeptide hydrogen bonds // Biophysics . 1983. V. 28(6). P. 962–965. (In Russ.)
3. Mazur A.K., Abagyan R.A. New methodology for computer-aided modelling of biomolecular structure and dynamics. 1. Non-cyclic structures // J. Biomol. Struct. Dyn., 1989. V. 6. P. 815–832.
4. Abagyan R.A., Mazur A.K. New methodology for computer-aided modelling of biomolecular structure and dynamics. 2. Local deformations and cycles // J. Biomol. Struct. Dyn, 1989. V. 6. P. 833–845.
5. Kramer R.Z., Bella J., Mayville P., Brodsky B., Berman H.M. Sequence dependant conformational variations of collagen triple-helical structure // J. Nat. Struct. Biol. 1999. V. 6. P. 454.
6. Berisio R., Vitagliano L., Mazzarella L., Zagari A. Crystal structure of the collagen triple helix model $[(pro\text{-}pro\text{-}gly)_{10}]_3$ // J. Protein Sci. 2002. V. 11. P. 262.
7. Benedetti E., Morelli G., Nemethy G., Scheraga H.A. Statistical and energetic analysis of sidechain conformations in oligopeptides // J. Pept. Protein Res. 1983. V. 22. P. 1–15.

SEARCHING STRUCTURAL PROTEIN DATABASES FOR ENZYMATIC ACTIVE SITES BY 3D PATTERNS

Gariev I.A.*, Uporov I.V., Varfolomeev S.D.

Moscow State University, Moscow, Russia
* Corresponding author: e-mail: gariev@hotmail.com

Keywords: *protein structure analysis, functional site, 3D-patterns*

Summary

Motivation: The current exponential growth of protein structural databases requires development of automated methods for search, analysis and assessment of protein function, such as methods that currently exists for protein sequence analysis.

Results: This work presents an accurate and automated method of search for the catalytic sites in structural databases and discriminating them from non-catalytic ones.

Methods

The search method is based on two premises: 1) residues of interest must be conserved among set of homologous proteins and 2) to be catalytic active they must be finely positioned in space.

It was shown that the methods based on geometrical constraints on atoms' level are valuable for search for proteins' functional sites, including active sites of enzymes (Wallace *et al*., 1997; Jones, Thornton, 2004).

The selectivity of the method can be increased by imposing conservativeness criterium filter on residues before applying geometrical-based search technique.

As a quantitative measure of residue conservativeness we use Shannon entropy, calculated for protein sequences with known structure in the HSSP database (Sander, Schneider, 1991).

We propose a simple language for defining 3D templates that allows user to impose arbitrary constraints on type of sought-for residues and geometrical constraints on position of their atoms.

Templates of catalytic sites of serine and cysteine hydrolases and acid proteases were derived by manual examination of several structures of each kind.

Results

A thorough analysis of sensitivity and selectivity of the method was made on non-redundant set of structures of enzymes with hydrolytic activity. Results were compared with information found in SwissProt database. Comparison shows that the method is highly accurate and can be used for database automatic annotation, for validation of information about known proteins and for automated detection of catalytic center residues in new structures. Moreover, the obtained results can be used for refinement of primary sequence templates, widely used for assigning protein function by primary sequence.

The method can quickly detect disagreements between residues found by protein sequence analysis and the actual catalytic ones in datasets of whole protein database. These disagreements shed light on protein evolution, when new enzyme appears from old by subtle changes in primary sequence.

Due to simple definition of sought-for site, the method can easily be extended to non-enzymatic functional sites of proteins.

The program (reference implementation of search method) is written by author (I.A.G.) in Perl.

References

Jones S., Thornton J.M. Searching for functional sites in protein structures // Curr. Opin. Chem. Biol. 2004. V. 8(1). P. 3–7.

Sander C., Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment // Proteins. 1991. V. 9(1). P. 56–68.

Wallace A.C., Borkakoti N., Thornton J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites // Protein Sci. 1997. V. 6(11). P. 2308–23.

A SYSTEM FOR COMPLEX ANALYSIS OF PROTEIN MACROMOLECULES SPATIAL STRUCTURES

Gribkov M.A.*, Korotkova M.A.

Moscow Engineering-Physics Institute (State University)
* Corresponding author: e-mail: lone_strider@mtu-net.ru

Keywords: *DNA-protein complexes, spatial structures, alignments, multiple spatial alignments, geometrical core of a family, subfamily detection, clusterization, detection of maximally complete sub-graphs*

Summary

Motivation: 3D structures of proteins, DNA and multimolecular complexes stores very important, and sometimes unique, information about biological macromolecules processes they are involved in. One of the major methods to reveal such information is a comparative analysis of families of closely alike structures. Fast growing of new structural information calls for development of new software utilities to perform on such analysis.

Results: in an effort to carry out such an analysis of structures families we have developed an algorithm and a program for detection of a family geometrical core and for automatic dividing family into a set of subfamilies.

Availability: http://www.dpidb.belozersky.msu.ru:8080/CoreProc/index.htm (not permanent, available by request).

Introduction

One of a problems is a modern bioinformatics is a classification of a proteins 3D structures and 3D structures of their complexes with DNA, RNA and ligands. The importance of this problem is conditioned by the fact that similarly spatially formed proteins often carry out similar functions.

Classifications on a basis of 3D structures likeness of separated protein domains lies at the heart of CATH (Orengo *et al*., 1997) and SCOP (Murzin *et al*., 1995; Lo Conte *et al*., 2002) databases, it is used by Dali program to fill FSSP (Holm, Sander, 1994–1997). New proteins spatial structures data comes with an increasing speed, and relating them to one of families on a basis of a spatial likeness is useful in many cases. Before putting a structure to one of the structures families it is necessary to form the families theirselves. There are a number of approaches to this problem (see references).

One of basic concepts for this kind of researches is *geometrical core* (Altman, Gerstein, 1995; Altman *et al*., 1997). Geometrical core is a subset of atoms, which spatial arrangement is conservative for all the structures in the family. For determining of a geometrical core algorithmically it is necessary to have a *multiple spatial alignmen*t of the family. Determining a spatial alignment of proteins structures family means determining an overall indexation of C-alpha atoms for all proteins in such a way, that atoms with identical indices have a similar spatial arrangement (maximally closely). The exact criterions of arrangement similarity differ for different authors.

Methods and Algorithms

The general steps of the algorithm is shown on the Figure 1. The role of input data for the algorithm is played by the multiple alignment of the structures (proteins or protein-DNA complexes) family (alignment itself could be represented as a user-prepared script or it could be constructed dynamically with the alignment constructor at the program preparations step – preparations step is not described

here because of its algorithmic triviality; this step it is a pure technical one). It is possible to show that for the geometrical core described above the process of its detection can be reduced to the detection of a maximally complete sub-graph (clique). The concerned graph is built in such a way that every its vertex corresponds to a one position in alignment, and two vertices are connected by an edge if they are conservative (the distance between them is almost the same for all structures in the family, i.e. two positions are conservative in the family if for any two structures in the family distances between those positions differs no more then on D, where D is a user-configurable value). To solve the problem of the clique detection we have developed a special algorithm for detecting an approximate solution. This algorithm is expected to complete the solution in a polynomial time.

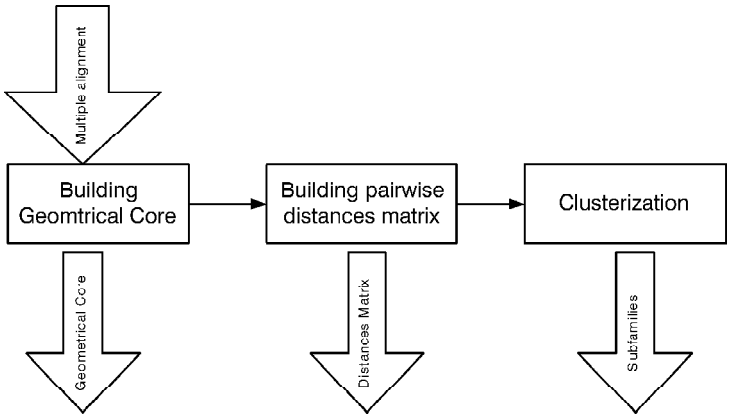


Fig. 1. General steps. At first we are building the main family geometrical core. Then, on its basis we're building the pairwise distances matrix. Distances matrix is used while clustering the family.

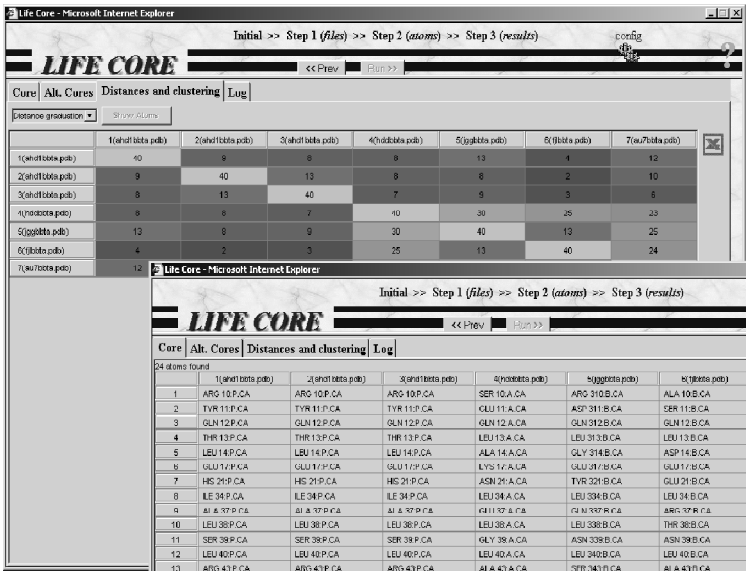


Fig. 2. LifeCore system.

Geometrical core is already very interesting itself, but moreover it is used to build a pairwise distances matrix (pairwise likeness matrix) for the family structures. Every cell of such matrix corresponds to a pair of structures and contains a number of atoms that could be added to the family geometrical core if the family would consist of only those two structures. Thus, the value of a matrix cell corresponding to a pair of structures characterizes their likeness degree.

For the final calculations stage (family clusterization into a set of subfamilies) we have developed a special algorithm to detect subfamilies on the principle of a subfamily geometrical core maximization. In context of a possible following comparison of new structures with the structures of the families detected, geometrical cores of those families could be counted as their "patterns".

Results and Discussion

The method described above was realized in the form of the LifeCore web-service. Figure 2 shows appearance of most important LifeCore windows.

LifeCore works with the PDB data bank copy and helps to analyze any families of existing structures by the user's choice. Calculation results (family geometrical core, pairwise distances matrix and a set of subfamilies) are shown to the user in a form of Java-applets graphical components and MS Excel tables.

Test runs for the LifeCore system were performed on several families such as "HTH3+POU" family and different complexes of DNA with homeodomains. Those runs shows results almost identical (with good precision) to the expected ones. It allows us to characterize the system developed as a quite appropriate one for the problems of protein macromolecules spatial structures relations analysis.

Acknowledgements

This work was partly supported by RFBR grant 03-07-90157.

References

Altman R.B., Schmidt B., Gerstein M. LPFC: a library of protein family core structures // Protein Science. 1997. V. 6(1). P. 246–248.

Altman R.B., Gerstein M. Using a measure of structural variation to define a core for the globins // Comput. Appl. Biosci. 1995. V. 11(6). P. 633–644.

Escalier V. Pairwise and multiple identification of three-dimensional common substructures in proteins // J. of Computational Biol. 1998. V. 5(1). P. 41–56.

Holm L., Sander C. The FSSP database of structurally aligned protein fold families // Nucleic Acids Res. 1994. V. 22(17). P. 3600–3609.

Holm L., Sander C. Dali: a network tool for protein structure comparison // Trends Biochem. Sci. 1995. V. 20(11). P. 478–480.

Holm L., Sander C. The FSSP database: fold classification based on structure-structure alignment of proteins // Nucleic Acids Res. 1996. V. 24(1). P. 206–209.

Holm L., Sander C. Dali/FSSP classification of three-dimensional protein folds // Nucleic Acids Res. 1997. V. 25(1). P. 231–234.

Leibowitz N., Fligelman Z., Nussinov R., Wolfson H. Automated multiple structure alignment and detection of a common substructural motif // Proteins: Structure, Function, and Genetics. 2001. V. 43. P. 235–345.

Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. SCOP database in 2002: refinements accommodate structural genomics // Nucl. Acid Res. 2002. V. 30(1). P. 264–267.

Murzin A.G., Brenner S.E., Hubbard T., Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures // J. Mol. Biol. 1995. V. 247. P. 536–540.

Orengo C.A., Michie A.D., Jones S, Jones D.T., Swindells M.B., Thornton J.M. CATH – a hierarchic classification of protein domain structures // Structure. 1997. V. 5(8). P. 1093–1108.

PDBSITE, PDBLIGAND AND PDBSITESCAN: A COMPUTATIONAL WORKBENCH FOR THE RECOGNITION OF THE STRUCTURAL AND FUNCTIONAL DETERMINANTS IN PROTEIN TERTIARY STRUCTURES COMBINED WITH PROTEIN DRAFT DOCKING

***Ivanisenko V.A.*[1]*, *Pintus S.S.*[2], *Krestyanova M.A.*[2], *Demenkov P.S.*[2], *Znobisheva E.K.*[2], *Ivanov E.E.*[2], *Grigorovich D.A.*[1]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Novosibirsk State University, Novosibirsk, Russia

* Corresponding author: e-mail: salix@bionet.nsc.ru

Keywords: *protein functional sites, site recognition, protein tertiary structure, draft docking, transcription factors classification, hepatitis C virus*

Summary

Motivation: The recognition of the structural/functional determinants in proteins has broad implications for structural genomics. A better understanding of the structural/functional determinants of proteins, such as protein-protein, protein-DNA, and protein-RNA interaction sites would provide insight into protein functions.

Results: A computational workbench for recognizing functional sites in protein tertiary structure combined with molecular draft docking was developed. Here, we illustrate the capabilities of the workbench by providing examples of search for interactions of the hepatitis C virus proteins with the human proteins and also of structural protein classification based on the structural similarity to the functional sites. The program PDBSiteScan is available at http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html. The PDBSite database is available at http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+PDBSite.

Introduction

Experimental data on protein tertiary structure are growing at a rapid pace (Westbrook *et al*., 2003). The body of literature is extensive. With the advent of methodologies for the recognition of functional sites in primary structure (Bairoch, Bucher, 1994), tools for site recognition in tertiary structure based on structural data on it alone (Ondrechen *et al*., 2001; Liang *et al*., 2003; Gutteridge *et al*., 2003), as well as on structural similarity to related proteins of known function (Wallace *et al*., 1997; Fetrow, Skolnick, 1998; Jones *et al*., 2003) were developed. There is now a repertoire of tools for the search of functional sites using databases containing structural data on protein-ligand interactions (Hendlich, 2003).

Research increasingly focuses on proteomics in efforts to clarify how ligand-protein binding sites may be recognized and to generate their complexes. Thus, the concept of molecular docking became popular (for an overview, see Schneidman-Duhovny *et al*., 2004).

We have developed the PDBSite database for the spatial structures of the protein functional sites, including the posttranslational modification and binding sites, the active enzyme centers. The created PDBSiteScan program provides search on the PDBSite database using pairwise protein-site structure alignment. Good recognition accuracy of the functional sites by screening of protein tertiary structure on the PDBSite database has been illustrated by the active enzyme centers (Ivanisenko *et al*., 2004).

Here, we extend and improve PDBSite by developing a PDBLigand database and a draft molecular docking module to further combine them. The draft docking module can help to solve an important

aspect of the molecular docking problem, the initial disposition of interacting molecules with respect to each other in space.

An approach to automated structural and functional classification of proteins on the basis of their structural similarity to the functional sites from the PDBSite database is proposed. The resulting classification of the representatives of the main transcription factor families agreed well with the standard manual classification (Heinemeyer *et al*., 1999).

An example is provided to illustrate the benefits of combining PDBSite + PDBSiteScan + PDBLigand into a common workbench, namely evidence for the possible role of RNA-directed RNA polymerase (NS5B) hepatitis C virus (HCV) in the regulation of host immunity.

Methods and Algorithms

The workbench consists of the PDBSite and the PDBLigand databases, PDBSiteScan program, and the draft docking module. The PDBSite database stores the data for the functional protein sites, the PDBLigand database those for the ligands of the sites.

A brief description of the PDBSite structure and the PDBSiteScan program follows (for details, see Ivanisenko *et al*., 2004; Ivanisenko *et al*., 2002). PDBSite contains more than 8,000 sites, including catalytically active centers of various enzymes, the sites of posttranslational protein modification, the sites of ion metal binding, the sites of binding organic/inorganic compounds, the sites of drug binding, the sites of protein-protein, protein-DNA and protein-RNA interactions. The data extracted from the PDB databank (Berman *et al*., 2000) on the basis of information in the SITE field of PDB indicating the amino acid residues of the functional sites; the sites of protein-protein, protein-DNA and protein-RNA interactions were identified by analysis of the atom coordinates in their heterocomplexes. The sites included the amino acid residues that are in contact with the ligand (protein, RNA or DNA). A residue was accepted as contact if it had at least three atoms whose distance from any atom of the partner chain was smaller than 5 Å .

The PDBLigand database contains data on the low molecular ligands, proteins, DNA and RNA, which bind to the sites from PDBSite. The PDBLigand database includes the atom coordinates of the ligands, also their functional description extracted from the PDB databank. Every entry of the PDBLigand database contains information on a particular ligand links to an entry of the PDBSite database providing information on the binding site of the ligand.

The PDBSite database is integrated with the PDBSiteScan program for recognizing the functional sites in protein tertiary structures. PDBSiteScan provides automated search of the spatial fragments in protein tertiary structure similar in structure to the functional sites from the PDBSite database.

The draft docking module works as follows. The PDBSite database contains the site-templates with known atom coordinates of their complexes with the ligands from the PDBLigand database. Draft docking is done by transfer of the ligand together with the site-template during the structural alignment of the site-template to protein. The generated draft protein-ligand complex can be accepted as a start approximation for the further docking or molecular dynamics analysis.

Implementation and Results

Search for the potential interactions between HCV and human proteins. The RNA dependent RNA polymerase NS5B is a 65 kDa protein that resembles other viral RNA polymerases (Lohmann *et al*., 1997). HCV replication is thought to occur in membrane bound replication complexes. The complexes transcribe the positive strand and the resulting minus strand is used as a template for the synthesis of genomic RNA. Search on the PDBSite database using PDBSiteScan demonstrated that NS5B contains fragments structurally similar to the binding site to the human nuclear transport factor 2 (NTF2) and to the human nuclear factor of activated T cells (NFAT). NTF2, a homodimer of approximately 14 kDa subunits, stimulates efficient nuclear import of a cargo protein (Stewart, 2000).

NFAT transcription factor family is involved in the expression of the cytokines IL-2, IL-3, IL-4, IL-5, granulocyte-macrophage colony-stimulating factor, and tumor necrosis factor-alpha, as well as several cell-surface molecules, such as CD40L and FasL. NFAT proteins are also expressed in B cells, mast cells, basophils and natural killer cells, as well as in a variety of non-immune cell types and tissues, such as skeletal muscle, neurons, heart and adipocytes (Porter *et al*., 2000).

The potential complexes of NS5B with NTF2 and NFAT generated using the draft docking module are shown in Fig. 1. It is seen that two loops are involved in the interaction of NS5B with NTF2 (Fig. 1a) and, hence, the contact might be close. Further calculations in terms of molecular dynamics, for example, are required to estimate contact affinity. The second complex results from contacts between only four residues at each site (Fig. 1b). However, NS5B can contact with the DNA bound to NFAT. It is suggested that the double contact of NS5B with NFAT and DNA can establish a stable complex. Molecular modeling is required to prove this.

Classification of transcription factors. Seventeen families of transcription factors were chosen for classification. Structural similarity to the functional sites from the PDBSite database were searched for every protein. The total number of functional site types examined was 88. The maximum distance mismatch (MDM) and amino acid type match were calculated to express the similarity between a protein fragment and a site (see Ivanisenko *et al*., 2004). A site and a protein fragment were accepted as structurally similar if the MDM value was less then 2 Å . The fragments structurally similar to the sites were further divided into four classes: 1) completely matching the amino acids; 2) one mismatch; 3) two mismatches and 4) three or more mismatches.

The distance between a pair of protein tertiary structures was calculated from $D_{ij} = \sqrt{\sum_{k=1}^{88}\sum_{m=1}^{4}(x_{km}^{i} - x_{km}^{j})^2}$, where i, j are the indices of protein tertiary structure, x_{km}^{i} is the variable indicating whether or not at least one fragment, structurally similar to a functional site of the k-th type and assigned to the m mismatch class, is present in the *i-th* protein structure; the values assigned to were either 1 or 0; 1 was assigned to a particular site type if at least one fragment in the protein structure was found to be similar to at least one site from PDBSite of this type; otherwise 0 was assigned to this site type. The UPGMA method was used for clustering, and the PHYLIP package (Lim, Zhang, 1999) to construct the hierarchical tree (Fig. 2).

Discussion

The developed workbench was designed for addressing problems related to the functional annotation and draft docking of proteins with low molecular weight ligands, proteins, RNA and DNA. The workbench was applied to the analysis of interactions of the HCV-human proteins. As a result, we identified the potential binding site NS5B of HCV to the human nuclear transport factor 2 (NTF2) and also to the DNA binding domain of the human transcription factor NFAT. The results suggest that the NS5B-NTF2 interaction provides NS5B transport into the cell nucleas where it interacts with NFAT and DNA, participates in the regulation of gene expression, thereby suppressing antiviral immunity. It should be noted that the assumption requires support: modeling of the NS5B-NTF2, NS5B-DNA-NFAT draft complexes.

Although the structural classification of proteins is a powerful clue to problems in proteomics, there are no universal algorithms. Structure alignment methods are difficult to implement because of the vagueness of their global similarity measures. The structure alignment methods often measure similarity by the root-mean-square-deviation (RMSD) between the aligned atoms. Rogen and Fain have indicated that the RMSD of aligned atom coordinates is a perfect measure of similarity for two shapes that are nearly identical (Rogen, Fain, 2003). However, the RMSD is a poor measure when the two shapes compared differ significantly. As a result, automated classification of proteins

remains an open issue. We suggested an approach to automated protein classification based on search for structural similarity between protein fragments and functional sites from the PDBSite database. The approach was applied to the classification of the representatives of the main classes of transcription factors. The resulting classification agrees well with the one obtained by manual analysis.

The proposed workbench has already proven itself to be useful in analysis. Further integration with other computational tools is possible and beneficial.

Fig. 1. The potential complexes of NS5B with NTF2 (a) and NFAT (b). The NS5B structure is dark grey, the NTF2 and NFAT structures are light grey. In the NS5B-NFAT complex, a fragment of double stranded DNA, which interacts with NFAT and presumably with NS5B, is depicted. The atom coordinates of NS5B, NTF2 and NFAT in complex with DNA were extracted from PDB 1QUV, 1A2K and 1A02, respectively.

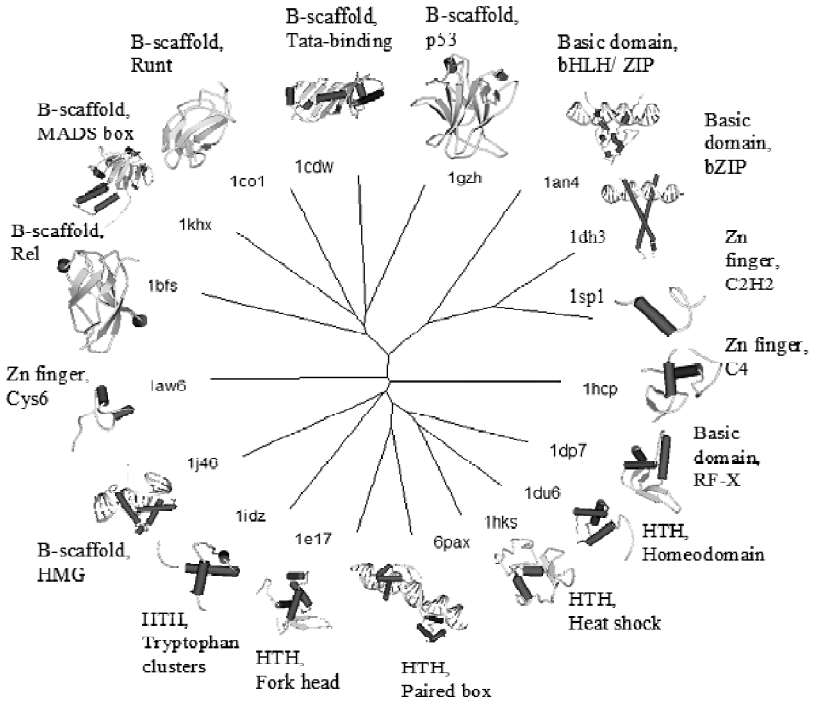


Fig. 2. A hierarchical tree for a classification of the representatives of the main classes of transcription factors. The tree was built on the basis of search for the structural homology of the DNA-binding domains of these factors with the functional sites from the PDBSite database. The name of the class, the PDB ID and a schematic representation of tertiary structure are given for every domain.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (02-07-90355, 03-07-96833-p2003, 03-07-96833, 03-07-90181-в, 02-04-48802-a, 03-04-48829, 03-04-48555-a); Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology"; the Siberian Branch of the Russian Academy of Sciences (integration project No. 119); Russian Ministry of Industry, Science, and Technologies (grants Nos. 43.073.1.1.1501); the U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF) No. NO-008-X1.

References

Bairoch A., Bucher P. PROSITE: recent developments // Nucleic Acids Res. 1994. V. 22. P. 3583–3589.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank // Nucleic Acids Res. 2000. V. 28. P. 235–242.

Fetrow J.S., Skolnick J. Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases // J. Mol. Biol. 1998. V. 281. P. 949–68.

Gutteridge A., Bartlett G.J. Thornton J.M. Using a neural network and spatial clustering to predict the location of active sites in enzymes // J. Mol. Biol. 2003. V. 330. P. 719–734.

Heinemeyer T., Chen X., Karas H., Kel A.E., Kel O.V., Liebich I., Meinhardt T., Reuter I., Schacherer F., Wingender E. Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms // Nucleic Acids. Res. 1999. V. 27. P. 318–322.

Hendlich M., Bergner A., Gunther J., Klebe G. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions // J. Mol. Biol. 2003. V. 326. P. 607–620.

Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A. PDBSite: a database on protein active sites and their environment // Third International Conference on Bioinformatics of Genome Regulation and Structure. 2002. V. 3. P. 145–148.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins // Nucleic Acids Res. 2004. V. 32, in press.

Jones S., Barker J.A., Nobeli I., Thornton J.M. Using structural motif templates to identify proteins with DNA binding function // Nucleic Acids Res. 2003. V. 31. P. 2811–2823

Liang M.P., Banatao D.R., Klein T.E., Brutlag D.L., Altman R.B. WebFEATURE: An interactive web tool for identifying and visualizing functional sites on macromolecular structures // Nucleic Acids Res. 2003. V. 31. P. 3324–3327.

Lim A., Zhang L. WebPHYLIP: A Web Interface to PHYLIP // Bioinformatics, 1999. V. 15. P. 1068–1069.

Lohmann V., Korner F., Herian U., Bartenschlager R. Biochemical properties of hepatitis C virus NS5B RNA-dependent RNA polymerase and identification of amino acid sequence motifs essential for enzymatic activity // J. Virol. 1997. V. 71. P. 8416–8428.

Ondrechen M.J., Clifton J.G., Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 12473–12478.

Porter C.M., Havens M.A., Clipstone N.A. Identification of amino acid residues and protein kinases involved in the regulation of NFATc subcellular localization // J. Biol. Chem. 2000. V. 275. P. 3543–3551.

Røgen P., Fain B. Automatic classiûcation of protein structure by using Gauss integrals // Proc. Natl Acad. Sci. USA. 2003. V. 100. P. 119–124.

Schneidman-Duhovny D., Nussinov R., Wolfson H.J. Predicting molecular interactions in silico: II. Protein-protein and protein-drug docking // Curr Med. Chem. 2004. V. 11. P. 91–107.

Stewart M. Insights into the molecular mechanism of nuclear trafficking using nuclear transport factor 2 (NTF2) // Cell Struct. Funct. 2000. V. 25. P. 217–225.

Wallace A.C., Borkakoti N., Thornton J.M. TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites // Protein Sci. 1997. V. 6. P. 2308–2323.

Westbrook J., Feng Z., Chen L., Yang H., Berman H.M. The protein data bank and structural genomics // Nucleic Acids Res. 2003. V. 31. P. 489–491.

SDPPRED: A METHOD FOR PREDICTION OF AMINO ACID RESIDUES THAT DETERMINE DIFFERENCES IN FUNCTIONAL SPECIFICITY OF HOMOLOGOUS PROTEINS AND ITS APPLICATION TO THE MIP FAMILY OF MEMBRANE TRANSPORTERS

Kalinina O.V.*[1], Novichkov P.S.[1], Mironov A.A.[1,2], Gelfand M.S.[2,3], Rakhmaninova A.B.[1]

[1] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; [2] State Scientific Center GosNIIGenetika, Moscow, Russia; [3] Institute for Problems of Information Transmission RAS, Moscow, Russia

***** Corresponding author: e-mail: ok81@yandex.ru

Keywords: *protein specificity prediction, mutual information, Bernoulli cutoff*

Summary

Motivation: The increasing volume of genomic data opens new possibilities for the analysis of protein function.

Results: SDPpred (Specificity Determining Position prediction) is a tool for prediction of residues in protein sequences that determine the proteins' functional specificity. It is designed for the analysis of protein families, whose members have biochemically similar but not identical interaction partners (e.g., different substrates for a family of transporters). SDPpred predicts residues that could be responsible for the proteins' choice of their correct interaction partners. SDPpred does not require information about secondary or 3D structure of proteins.

Availability: SDPpred is available at http://math.belozersky.msu.ru/~psn/.

Introduction

Many protein families contain homologous proteins that have a common biological function, but different specificity towards substrates, ligands, effectors, DNA, proteins and other interacting molecules, including other monomers of the same protein. All these interactions must be highly specific. The proteins can be assigned to specificity groups based on experimental data or comparative genomic analysis.

Identification of residues that account for the protein specificity might be useful in many biological studies. One obvious application of SDPpred is to minimize the number of point mutations required to switch the specificity of an enzyme, regulator or transporter. Analysis of the predicted residues can also provide a deeper insight into the nature of functional specificity. Prediction of SDPs is also reasonable for families containing specificity groups of any nature (e.g., proteins of different thermostability).

Amino acid residues that determine differences in the protein specificity and account for correct recognition of interaction partners, are usually thought to correspond to those positions of a protein multiple alignment, where the distribution of amino acids is closely associated with grouping of proteins by specificity. SDPpred searches for positions that are conserved within specificity groups but differ between them. These positions are called *SDPs (Specificity-Determining Positions).* Such positions, though obvious in alignments containing a small number of proteins and specificity groups, become a challenge to find in large protein families with a variety of specificities.

Recently, a number of algorithms addressing the described problem have been developed (for a review of some of these methods see Hannenhalli, Russell, 2000; Mirny, Gelfand, 2002; Kalinina

et al., 2004). SDPpred implements the algorithm described in (Kalinina *et al.*, 2004). Compared with other methods, this algorithm has several advantages. Firstly, it does not use any information about the protein structure. The procedure is based solely on statistical analysis of an alignment. Secondly, it automatically calculates the number of SDPs and the probability of occurrence of these positions by chance. It does not incorporate any *ad hoc* cutoff setting. Thirdly, substitutions are weighted according to physical properties of amino acids, using a substitution matrix, so that substitutions to amino acids with similar properties are only weakly penalized. And finally, SDPpred incorporates information about evolutionary distance within and between groups by using different amino acid substitution matrices.

Algorithm and Web interface description

The only information needed for prediction of SDPs is a multiple alignment of protein sequences divided into specificity groups. We search for positions of a multiple protein alignment, for which the distribution of amino acid frequencies is closely associated with grouping by specificity. As a measure of such association we calculate *the mutual information* of each column *p* of the alignment:

$$I_p = \sum_{i=1}^{N} \sum_{\alpha=1}^{20} f_p(\alpha,i) log \frac{f_p(\alpha,i)}{f_p(\alpha) f(i)},$$

where $\alpha = 1,...,20$ is a residue type, $f_p(\alpha,i)$ is the ratio of the number of occurrences of residue α in group *i* at position *p* to the length of the alignment column, $f_p(\alpha)$ is the frequency of residue α in the alignment column, $f(i)$ is the fraction of proteins belonging to group *i*. High values of the mutual information indicate columns with high correlation between amino acid frequencies and the grouping by specificity.

To address the fact that the frequencies are calculated based on a small sample, and that substitutions to amino acids with similar physical properties should be weakly penalized, the observed amino acid frequencies are modified. Instead of using $f(a,i) = n(a,i) / n(i)$, where $n(\alpha,i)$ is the number of occurrences of residue α in group i, $n(i)$ is the size of group *i* (here *i* is a single group or the whole alignment), SDPpred uses *smoothed frequencies*

$$f'(\alpha,i) = \frac{n(\alpha,i) + \kappa \left(\sum_{\beta=1}^{20} n(\beta,i) m(\beta \to \alpha) \right) \Big/ \sqrt{n(i)}}{n(i) + \kappa \sqrt{n(i)}},$$

where $m(\beta \to \alpha)$ is the probability of amino acid substitution $\beta \to \alpha$ according to the matrix corresponding to the average identity in group *i*, $0 \le \kappa \le 1$ is a smoothing parameter. Additionally, the necessary pseudocounts are introduced in a natural way.

Then statistically significant values of the mutual information are selected using a novel procedure, called *the Bernoulli estimator* and is described in detail in (Kalinina *et al.*, 2004). Briefly, we search for those positions, which are least probable to be obtained by chance. These positions we call SDPs (Specificity Determining Positions).

SDPpred outputs the set of SDPs, i.e. positions of the alignment, which are likely to determine differences in the functional specificity among the given groups. This set can be visualized as an alignment of the family with the SDPs highlighted, a detailed description of each SDP, or a plot of probabilities, from which the minimum is chosen to set the cutoff. SDPpred is publicly available at http://math.belozersky.msu.ru/~psn/.

Results and Discussion

The results of testing, which agree well with available structural and experimental data, are described in (Kalinina *et al*., 2004). In that study, we analyzed two protein families: the LacI family of bacterial transcription factors and the MIP family of membrane channels in bacteria. Both these families include proteins with the resolved 3D structure, which was used to evaluate predictions. In both cases the fraction of contacting residues among SDPs is much larger than in the whole alignment (Table).

Table. Residues of different contact types among SDPs and in the whole alignment of the MIP and LacI protein families

	SDPs for the MIP family	Whole alignment of the MIP	SDPs for the LacI family	Whole alignment of the LacI
Contact (distance to an interaction partner <5Å)	13	95	22	82
Possible contact (distance to an interaction partner 5-10Å)	8	73	19	89
Not contact (distance to an interaction partner >10Å)	0	113	3	177
Total	21	281	44	348

In both cases the proteins function as oligomers and a substantial fraction of SDPs lies on the surface of contact between subunits. For example, in the case of the MIP family we predicted 21 SDPs, which were mapped onto the 3D structure of the tetramer of GlpF from *E. coli* (Fu *et al.*, 2000). Sixteen of them (22Ile, 48Trp, 135Phe, 136Ser, 137Thr, 159Leu, 187Ile, 191Gly, 194Met, 195Gly, 199Gly, 200Phe, 201Ala, 207Asp, 211Lys, 236Pro) either contact the GlpF substrate glycerol or lie in the channel-forming helices on the side exposed into the channel, whereas the remaining five (20Leu, 24Ile, 43Glu, 108Tyr, 193Ser) lie on the outer surface of the monomer (Kalinina *et al*., 2004) (Fig. 1) and contact other subunits (Fig. 2). These five SDPs form two types of spatial clusters: one A-type cluster formed by 43Glu of all four monomers of the GlpF tetramer, and four B-type clusters formed by 20Leu, 24Ile and 108Tyr of one subunit and 193Ser of another subunit (for details see (Kalinina *et al.*, in press).

This spatial arrangement of amino acid residues corresponding to SDPs suggests that evolutionary pressure on amino acids that establish intersubunit contacts is correlated with the evolutionary pressure on amino acids that account for the correct recognition of interaction partners. In the case of the MIP family, the residues lying on the surface of contact between subunits cluster together, possibly forming "structural clasps" that prevent formation of chimeric aquaporin-glyceroporin tetramers.

SDPpred can be applied to any protein family that includes proteins of different specificity. It produces results that agree with available structural and experimental data. It proved to be useful not only for identification of candidate sites for protein functional redesign or prediction of specificity of family members, for which the latter is unknown. It also provided deeper insight into the nature of protein-protein interactions and the mechanism of molecular recognition.

Acknowledgements

This study was partially supported by grants from HHMI (55000309), LICR (CRDF RB0-1268), RFBR (04-04-49440), the Fund for Support of Russian Science, and the Program in Molecular and Cellular Biology of the Russian Academy of Sciences.

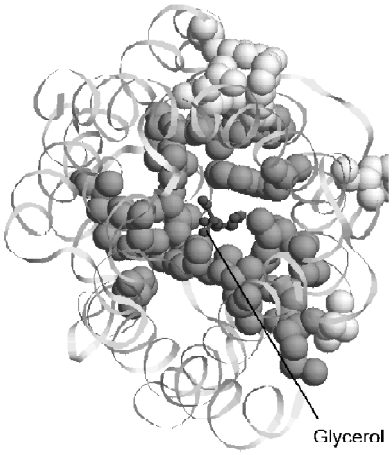


Fig. 1. Predicted SDPs mapped onto the structure of GlpF from *E. coli* (1fx8). Channel-forming SDPs are shown as gray spheres. SDPs located on the outer surface of the monomer are shown as white spheres.

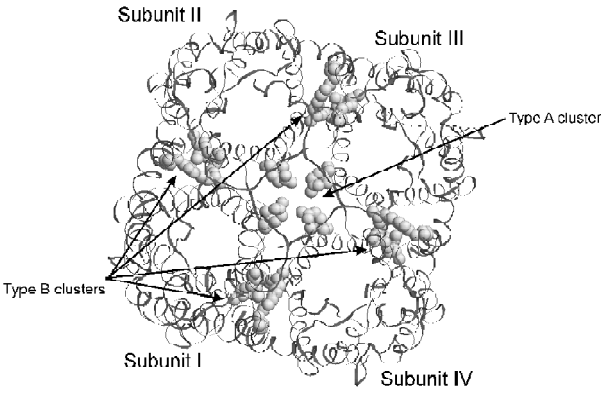


Fig. 2. Residues making "structural clasps" in the structure of the tetramer of the GlpF of *E. coli* (1fx8, biological subunit). SDPs lying on the surface of contact between subunits are shown as white spheres.

References

Fu D., Libson A., Miercke L.J., Weitzman C., Nollert P., Krucinski J., Stroud R.M. // Science. 2000. V. 290. P. 481–486.

Hannenhalli S.S., Russell R.B. Analysis and prediction of functional sub-types from protein sequence alignments // J. Mol. Biol. 2000. V. 303. P. 61–76.

Kalinina O.V., Mironov A.A., Gelfand M.S., Rakhmaninova A.B. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families // Protein Sci. 2004. V. 13. P. 443–456.

Kalinina O.V., Gelfand M.S., Mironov A.A., Rakhmaninova A.B. Amino acid residues forming specific contacts between subunits in tetramers of the membrane channel GlpF // Biofizika, (in press).

Mirny L.A., Gelfand M.S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors // J. Mol. Biol. 2002. V. 321. P. 7–20.

SYMMETRY AND SPATIAL STRUCTURE OF THE CANONICAL SET OF AMINO ACIDS

Karasev V.A.*[1], Luchinin V.V.[1], Stefanov V.E.[2]

[1] Saint-Petersburgh State Electrotechnical University "LETI", Saint-Petersburgh, Russia;
[2] Saint-Petersburgh State University, Saint-Petrsburgh, Russia
***** Corresponding author: e-mail: cmid@eltech.ru

Keywords: *canonical set of amino acids, genetic code, icosahedron, dodecahedron, spatial structure*

Summary

Motivation: The nature of the canonical set of twenty amino acids remains unsolved problem. In this connection we undertook analysis of group properties of the amino acids, which compose the canonical set. The dodecahedron structure was used for pictorial rendition of the derived principles.

Results: Analysis of the properties of a set of 12 meridian cycles obtained on the structure of the duplet genetic code, which is isomorphic to Boolean hypercube B^4, revealed four groups of cycles united in pairs by anti-symmetry transformations of two types. These transformations become most illustrative when shown on the icosahedron, a polyhedron with 12 vertices. Related to the icosahedron is another polyhedron – dodecahedron, which has 20 vertices. Approach based on the use of the two polyhedrons was applied to the analysis of structure of the canonical set of 20 amino acids. It was demonstrated that four groups of amino acids, each containing five amino acids connected by anti-symmetry transformations of two types, can be distinguished in the initial set. The revealed principles were pictorially represented on the structure of dodecahedron.

Availability: http://genetic-code.narod.ru/

Introduction

Development of spatial models of the duplet and triplet genetic code isomorphic to Boolean hypercubes B^4 and B^6, respectively (Klump, 1993; Jimenez-Montano *et al*., 1996; Karasev, Sorokin, 1997), is an important achievement. However, the proposed structures deal with the duplet and triplet code only, ignoring the nature of the canonical set of 20 amino acids. This set must have its structural principles, which up to now remain obscure. The genetic code should be regarded as a natural system of amino acid organization. Within the model developed by us (Karasev, 2003; Karasev, Stefanov, 2001), side chains of amino acids, encoded by triplets, are treated as physical operators reconstructing the encoded structure. They include connectivity operators (polar amino acids), encoded by codons, which have G or A in the second position of the triplet, and anti-connectivity operators, encoded by triplets with C or U in the second position. Another classification, based on the genetic code, addresses the idea of complementarity of amino acids encoded by complementary triplets (Mekler, Idlis, 1993).

There are alternative approaches based on the classification of the amino acid side chains according to their physicochemical properties (Campbell, Smith, 1994) and on the particular character of functioning of amino acids in the protein structure (Karasev, 2003).

However, the system of twenty amino acids may have its own spatial representation only indirectly connected with the code. The present study aims at the analysis of the structure of the canonical set of amino acids, based on principles of symmetry and anti-symmetry, and development of a spatial model, which would provide an illustrative representation of the structure of the set. Earlier this problem was addressed in a preliminary study (Karasev, 2004).

Model

The following prerequisites of the model should be indicated: a) coincidence of the number of amino acids in the canonical set with that of vertexes in the dodecahedron; b) occurrence of a structure related to dodecahedron, i.e. icosahedron; c) occurrence of functional connection of the side chains of amino acids with the spatial-temporal self-organization of protein molecules; d) possibility to use spatial structures, isomorphic to Boolean hypercubes, for describing the process of self-organization of protein molecules; e) establishing of the fact that the icosahedron represents the spatial structure of the meridian cycles of the duplet genetic code.

The structure of the duplet genetic code is known to be isomorphic to the Boolean hupercube B^4 (Karasev, 2003). Let us move on the structure of the genetic code (Fig. 1a) from the vertex situated in the first tier to the vertex situated in the fifth tier and then back along the symmetrical path. The resulting cyclic pathway, shown in (Fig. 1b), can be called meridian cycle (M-cycle).

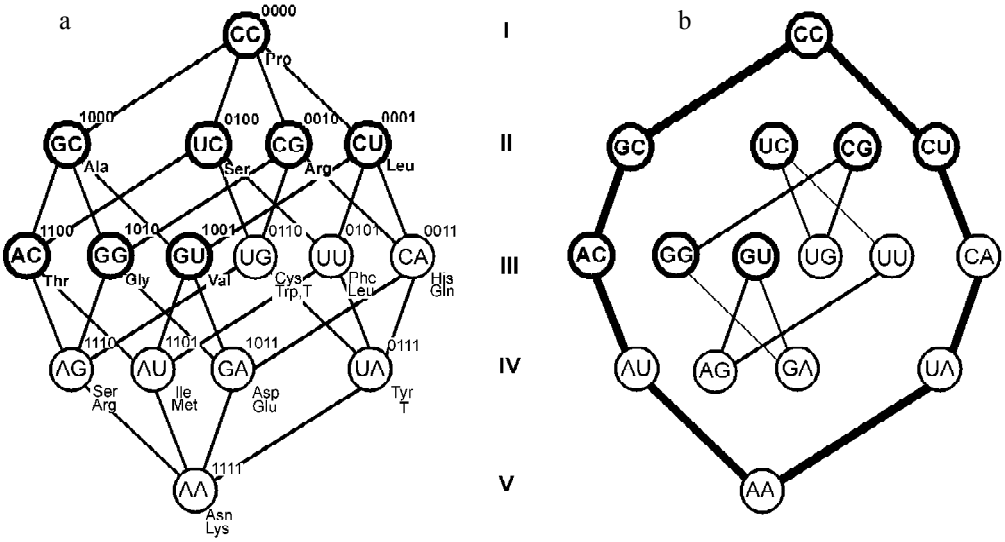


Fig. 1. Structure of the duplet genetic code (a) and meridian cycle highlighted on this structure (b).

The total number of M-cycles connected by relations of anti-symmetry, which can be identified on the structure of the duplet genetic code, is equal to twelve. Icosahedron is the spatial structure, which most obviously incorporates anti-symmetry principles of M-cycles (Fig. 2).

As seen from Fig. 2, the plane I separates M-cycles, interrelated by operation of anti-symmetry implying that letters of the duplets interchange according to the following rule: C←→A, G←→U. Plane II separates two groups of M-cycles, which do not have duplets in common other than CC and AA, the so called antipode cycles. On rotating this plane about the axis perpendicular to plane I (C_2), vertexes with antipode cycles coincide. Since icosahedron is related to dodecahedron, we applied the anti-symmetry principles established on the icosahedron to analyze the arrangement of amino acids on the dodecahedron.

Results and Discussion

The performed analysis revealed two groups of amino acids which have similar properties but different structure, e.g. Lys –Arg, Glu – Asp, Asn – Gln, two groups of amino acids with opposite properties, e.g. Lys – Glu, Arg – Asp, etc. These results are clearly demonstrated on the

dodecahedron structure (Fig. 3). Amino acids with similar properties but differing in size, occupy positions symmetrical with respect to plane I, whereas amino acids with opposite properties, occupy vertexes, which coincide with each other upon rotation of the dodecahedron about axis C_2. Thus, the set of 20 canonical amino acids consists of four groups, containing five amino acids each, which are interrelated by anti-symmetry transformations. At present, we carry out analysis of proteins on the basis of the developed structure model looking for construction principles for polypeptide chains that can be implemented in nanoelectronics and sensorics.

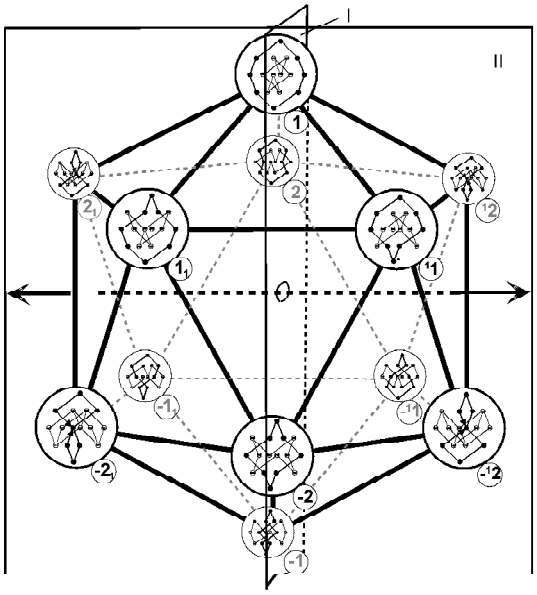


Fig. 2. Localization of meridian cycles of the duplet genetic code on the icosahedron. (I) – anti-symmetry plane separating antipodes of the 1st type; (II) – plane separating antipodes of the 2nd type.

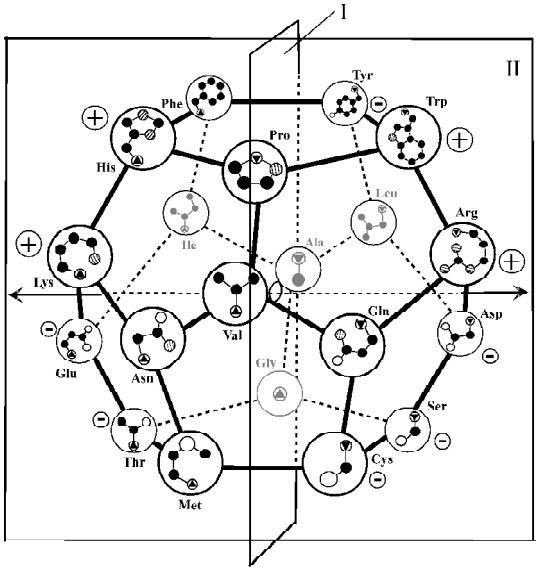


Fig. 3. System of amino acids, connected by relations of anti-symmetry, constructed on the dodecahedron. (I) – anti-symmetry plane separating amino acids with similar properties; (II) – plane separating amino acids with opposite properties.

References

Campbell P.N., Smith A.D. Biochemistry Illustrated. Edinburgh – London – Madrid – Tokio: Curchill Livingstone, 1994. P. 8–9.

Jimenez-Montaño M.A., de la Mora-Basañez C.R., Poschel Th. The hypercube structure of the genetic code explains conservative and non- conservative aminoacid substitutions *in vivo* and *in vitro* // BioSystems. 1996. V. 39. P. 117–125.

Karasev V.A. Genetic Code: New Horizons. SPb: Tessa, 2003. 145 p. (Russ.).

Karasev V.A. 2004. On anti-symmetry of the canonical set of amino acids. Dep.VINITI, 23.03.2004, N 470-B2004 (Russ.).

Karasev V.A., Sorokin S.G. Topological structure of the genetic code // Russ. J. Genetics. 1997. V. 33. P. 622–628.

Karasev V.A., Stefanov V.E. Topological nature of the genetic code // Theor. Biol. 2001. V. 209. P. 303–317.

Klump H.H. The physical basis of the genetic code: the choice between speed and precision // Arch. Biochem. Biophys. 1993. V. 301. P. 207–209.

Mekler L.B., Idlis R.G. General stereochemical genetic code – towards biology and universal medicine of the XXI century // Priroda. 1993. N 5. P. 29–63 (Russ.).

STATISTICAL METRICS FOR PROTEIN ACTIVE SITE PREDICTION WITH THEMATICS

***Ko J.*[1], *Andre P.*[2], *Murga L.F.*[2], *Ondrechen M.J.**[2]

[1] NSF-ROA awardee on leave from Department of Chemistry, Indiana University of Pennsylvania, 975 Oakland Avenue, Indiana, Pennsylvania 15705 USA; [2] Department of Chemistry and Chemical Biology and Institute for Complex Scientific Software, Northeastern University, Boston, Massachusetts 02115 USA
* Corresponding author: e-mail: mjo@neu.edu

Keywords: *active site, functional genomics, THEMATICS, titration*

Summary

Motivation: With genome sequences available for 10^3 species and structure determination of thousands of proteins underway, function prediction is the next major step. Computational methods are critical to these functional genomics efforts.

Results: THEMATICS is a simple computational method for the prediction of a protein's active site from its three-dimensional structure alone. The method identifies clusters of residues in physical proximity with perturbed predicted titration behavior. Originally the identification of the perturbed residues was done by simple observation of the plotted titration curves. In order to use the method for high-throughput screening, we have developed statistical metrics to classify the residues as either ordinary or perturbed and thus bypass the time-consuming visual observation step. Here we present a very simple statistical test that identifies the perturbed residues with high accuracy (about 93 %).

Introduction

Structural genomics projects are rapidly increasing the number of known three-dimensional structures of gene products, but these new structures generally are of unknown function. The determination of the function from the structure has proved to be more difficult than originally presumed. THEMATICS (Ondrechen, 2001; Shehadi, 2002; Ringe, 2004) is a new technique based on established Poisson-Boltzmann methods and it reliably predicts information about the active site of a protein from its three-dimensional structure alone. This method does not depend on any sequence analysis.

Model

THEMATICS is based on Poisson-Boltzmann methods for the calculation of the electrical potential function of a protein structure, followed by a Monte Carlo procedure to compute the proton occupations of the ionizable sites as a function of the pH (Bashford, Karplus, 1991; Gilson, 1993; Yang, 1993; Sampogna, Honig, 1994; Karshikoff, 1995; Antosiewicz, 1996). A small fraction of the residues in proteins are predicted to exhibit perturbed, non-Henderson-Hasselbalch titration behavior. We have argued that such residues with perturbed titration behavior tend to occur in the active site of an enzyme with sufficient frequency such that they serve as useful markers of chemical reactivity. THEMATICS identifies these perturbed residues and searches for clusters of two or more such residues in physical proximity, a reliable predictor of the location of the active site of an enzyme.

A typical ionizable residue in a protein obeys the Henderson-Hasselbalch equation, which may be written to express the mean net charge C on the residue as a function of x, where x is the difference between the pH and the pK_a ($x = pH - pK_a$), as:

$$C(x) = \pm (10^{\pm x} + 1)^{-1}. \quad (1)$$

In equation (1) the positive sign applies to residues that form a cation upon protonation and the negative sign applies to residues that form an anion upon deprotonation.

Figure 1 shows the predicted titration curves, net charge as a function of pH, for selected cysteine and tyrosine residues in the enzyme glutamate racemase from *Aquifex pyrophilus*. Note the nonsigmoidal shape of the curves for the catalytic residues C70 and C178, compared with the more typical sigmoidal shape of the non-active site residue C54.

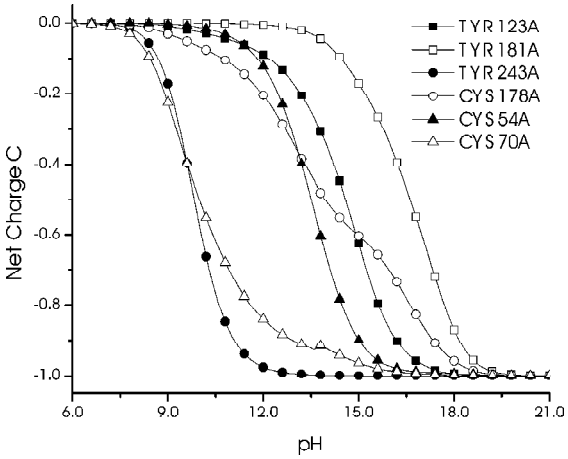


Fig. 1. Titration curves predicted for selected Cys and Tyr residues of glutamate racemase. Note the nonsigmoidal shape for the catalytic residues C70 and C178.

Originally a human observer would identify the titration curves that deviate from the typical sigmoidal shape. The deviant curves are typically identifiable by regions of shallow slope in the predicted titration curves. Thus, to automate the identification of the perturbed titration curves, it makes sense to examine their first derivative functions – dC/dpH. Figure 2 shows this first derivative as a function of pH for the same residues of glutamate racemase as shown in Figure 1. These functions resemble distribution functions and the area under each of these curves is unity. Hence we can define the nth central moment μ_n of the first derivative as:

$$\mu_n = \int_{-\infty}^{\infty} (pH - M_1)^n \left[\frac{dC}{d(pH)}\right] d(pH) \qquad (2)$$

where M_1 is the first raw moment, given by:

$$M_1 = \int_{-\infty}^{\infty} (pH) \left[\frac{dC}{d(pH)}\right] d(pH) . \qquad (3)$$

The Z score for the nth central moment is defined as the deviation from the mean value (averaged over all ionizable residues in the protein), divided by the standard deviation, as:

$$Z_n = (\mu_n - \langle \mu_n \rangle)\sigma_n \qquad (4)$$

For the odd-numbered moments, statistical analysis was performed on their absolute values. The third central moment μ_3, a measure of the asymmetry of the first derivative function, and the fourth central moment μ_4, a measure of the kurtosis or flatness of the function, have proved to be good metrics for the identification of deviant titration curves.

The following simple rule was used to classify predicted titration curves as non-sigmoidal: If either $Z_3 > 1.0$ or $Z_4 > 1.0$ for a given ionizable residue, that residue is designated as deviant and we label it as a THEMATICS positive residue. Clusters of two or more positive residues in physical proximity constitute a prediction of a chemically reactive site in the protein structure.

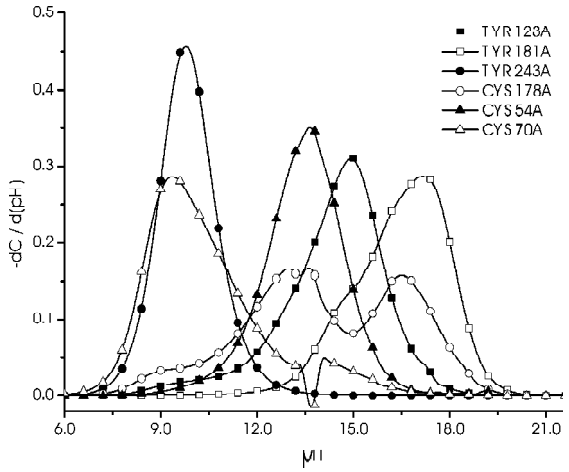


Fig. 2. First derivative -dC/dpH as a function of pH for the same residues of glutamate racemase as shown in Figure 1.

Results and Discussion

The statistical classifier does almost as well as the human observer and enables automated identification of active sites. The method was tried on 44 proteins that have already been well characterized experimentally. The Z_3 or Z_4 test identified the active site correctly for 41 of these proteins (93 %). Some representative results of the statistical test are tabulated, with known active site residues shown in boldface.

Enzyme	**Species**	**THEMATICS cluster prediction**
Glutamate racemase	*Aquifex pyrophilus*	[**D7, C70,** C139, **C178, E147,** E148, **H180,** Y39, Y123]
Deoxyribonuclease I	*Bos Taurus*	[D168, **D212, E39, E78, H134, H252**]
Beta lactamase	*E. coli*	[D133, D214, **E166, K73, K234**]

Acknowledgements

This work was supported by the National Science Foundation under grant MCB-0135303 and by the Institute for Complex Scientific Software at Northeastern University.

References

Antosiewicz J., McCammon J.A., Gilson M.K. The determinants of pKa's in proteins // Biochemistry. 1996. V. 35. P. 7819–7833.

Bashford D., Karplus M. Multiple-site titration curves of proteins: an analysis of exact and approximate methods for their calculation // J. Phys. Chem. 1991. V. 95. P. 9556–9561.

Gilson M.K. Multiple-site titration and molecular modeling: two rapid methods for computing energies and forces for ionizable groups in proteins // Proteins. 1993. V. 15(3). P. 266–82.

Karshikoff A. A simple algorithm for the calculation of multiple site titration curves // Protein Engineering. 1995. V. 8. P. 243–248.

Ondrechen M.J., Clifton J.G., Ringe D. THEMATICS: a simple computational predictor of enzyme function from structure // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 12473–12478.

Ringe D., Wei Y., Boino K.R., Ondrechen M.J. Protein structure to function: insights from computation // Cellular Mol. Life Sci. 2004. V. 61. P. 387–392.

Sampogna R.V., Honig B. Environmental effects on the protonation states of active site residues in bacteriorhodopsin // Biophys. J. 1994. V. 66(5). P. 1341–52.

Shehadi I.A., Yang H., Ondrechen M.J. Future directions in protein function prediction // Mol. Biol. Reports. 2002. V. 29. P. 329–335.

Yang A.S., Gunner M.R., Sampogna R., Sharp K., Honig B. On the calculation of pKas in proteins // Proteins. 1993. V. 15(3). P. 252–65.

FROM PROTEIN SEQUENCE TO PROTEIN SPECIFICITY: COMPLETELY AUTOMATED DISCOVERY AND MAPPING OF SPECIFICITY DETERMINING RESIDUES

Kolesov G.*, Mirny L.A.

Harvard-MIT Division of Health Sciences and Technology, MIT, Cambridge, MA 02139, USA
***** Corresponding author: e-mail: g.kolesov@mail.ru

Keywords: *protein structure, specificity determining residues, mutual information, evolution, phylogenetics, multiple sequence alignment*

Summary

Motivation: Most proteins bind and recognize specific targets such as other proteins, specific sites on DNA or RNA and small molecules. Specific recognition is essential for all cellular functions and is frequently affected in diseases. A detailed molecular understanding of protein specificity is essential for structure-based drug design, understanding of effects of mutations and rational manipulation of protein specificity.

Results: A robust, completely automated bionformatics framework for the recognition of specificity determining residues has been developed and tested.

Availability: At the time of this writing the program is not available for public use. Computations for individual proteins can be performed on request to the authors. We plan to make our method available as a World Wide Web server in the future.

Introduction

In the recent years a variety of approaches have been proposed aiming to exploit evolutionary relationships within a given protein family and pinpoint amino-acid residues that determine the specificity of the protein. These approaches often rely on additional information and/or additional expert analysis. For instance, reliable predictions of specificity determining residues (SDRs) often require the computation of phylogenetic trees, availability of high quality multiple sequence alignments (MSAs) and visual inspection of protein 3D-structure models (Lichtarge, Sowa, 2002; Mirny, Gelfand, 2002; Kalinina *et al.*, 2004).

Here we report a completely automated bioinformatics approach, which for a given protein sequence 1. finds members of the same family in the sequence database, 2. partitions the resulting set to orthologous groups, 3. computes the MSA, 4. locates putative SDRs (PSDRs), 5. estimates statistical significance of the PSDRs, 6. maps the significant positions on the protein's 3D-structure model (if available) and 7. generates a web report page.

Methods

Assembling and partitioning sequence set. When entire protein sequences are being used, the construction of reliable MSAs is hindered or simply infeasible due to the multi-domain structure of proteins and domain shuffling events in the course of evolution. A better approach is to consider individual domains rather than entire protein.

In the first step of the algorithm a PFAM (Bateman *et al*., 2000) domain search is conducted for the query sequence. Domains contributing to the protein are then analyzed separately. Alternatively, a PFAM hidden Markov model (HMM) domain profile can be fed directly into the program to initiate the analysis. We shall refer to a domain sequence as simply a 'sequence' further in this text.

Associated to a sequence PFAM HMM profile is then run against a protein sequence database to harvest homologous domains originating from different organisms. To find orthologous domains all-against-all BLAST (Altschul *et al*., 1997) alignment is then conducted for the resulting sequence set. Bi-directional best BLAST hits are identified.

Sequences are clustered into orthologous groups by using two approaches depending on the size of the set. If the set is small (<150 sequences) the single linkage clustering of best-to-best BLAST hits is applied.

In the cases when the set is large enough we select sequences originating in the same organism (usually the organism containing original sequence or an arbitrary one) as the 'parents' of the clusters. The 'parent's' best-to-best BLAST matches are included in the cluster. Overlaps are resolved by omitting the second and following occurrences of the same sequence in the clusters. Clusters containing a single sequence are removed. The underlying assumption of this method is that the organism contains only a single instance of each orthologous domain. While this is considered to be good estimate for prokaryotes with their compact genomes, in eukaryotic genomes the situation can be different and more sophisticated techniques such as INPARANOID (Remm *et al*., 2000) should be applied.

MSA. Multiple sequence alignment is conducted by either CLUSTALW (Thompson *et al.*, 1994) or the *hmmalign* program from the HMMER package (Eddy S.R., 1998) using default parameters.

We did not find any significant difference between MSAs obtained from these two programs.

Locating PSDRs with mutual information. To identify residues that can discriminate between paralogous proteins (different specificity), merging orthologs (same specificity) together, we used the mutual information as a measure of association with the specificity:

$$I_i = \sum_{x=1..20} \sum_{y=1..Y} P_i(x,y) \log \frac{P_i(x,y)}{P_i(x)P(y)}, \qquad (1)$$

where *i, x* and *y* denote the position in the alignment, the amino acid type and the orthologous group number (the same for all proteins of the same specificity group) respectively. *Y* is the total number of groups in an alignment. $P_i(x,y)$ is the joint probability of *x* and *y*, i.e. the probability of finding amino acid type *x* at position *i* and in group *y*. $P_i(x)$ is the marginal probability of finding amino acid type *x* at position *i* regardless of groups, and *P(y)* is simply the fraction of proteins belonging to group *y*. Importantly, I_i measures the correlation between *x* and y, and $I_i = 0$ if and only if *x* and y are statistically independent (Cover, Thomas, 1991).

Mutual information has several important properties: 1) it is non-negative; 2) it equals zero if and only if *x* and *y* are statistically independent; and 3) a large value of I_i indicates a strong association between *x* and *y*. Unfortunately, a small sample size and a biased amino acid composition of each column in the MSA influence I_i a lot. For example, positions with less conserved residues tend to have higher mutual information. Hence, we cannot rely on the value of I_i as an indicator of specificity association, instead we estimate the statistical significance of I_i.

Estimating statistical significance of PSDRs. To evaluate the statistical significance of an I_i, we need "control" MSAs to estimate the $P(I_i)$ and the *p*-value (the probability of observing this or higher I_i in the control). The control-MSA should carry most of the properties of the real MSA. We base our choice on the following reasoning. There are two major mechanisms of conservation for an amino acid position.

1. The first mechanism is independent of amino acid position in a protein. Examples of the first mechanism are chance and phylogeny (Wollenberg, Atchley, 2000). Position-independent signal can be take into account if we construct a "control" MSA using **the same sample size and the same phylogenetic tree**.

2. The second mechanism is position-specific, and conserves residues that have important structural

or functional role. For example, highly hydrophobic positions in MSA reflect selection of a residue in the hydrophobic core that stabilizes protein structure. A glutamate residue conserved across all proteins in the family can participate in the active site or in non-specific binding. In other words, position-dependent selection (other than sought diversifying selection of SDRs) is manifested in MSA columns with a biased amino acid composition. To compute statistical significance and rule out high I_i due to other evolutionary signals, we construct a "control"-MSA that has **the same amino acid composition** at every position i.

We developed two different methods to compute $P(I_i)$, which, however, produced very similar results. The first one takes into account only (2) and then compensates for the position-independent signal. The second one simulates protein evolution under selection and explicitly generates MSA that satisfies (1) and (2). Both methods are described in detail in (Mirny, Gelfand, 2002).

Mapping to a 3D-structure model. Mapping to a Protein Databank (PDB) protein 3D-structure model is achieved through the ability of the hmmalign program to align a sequence to the MSA (using the MSA's HMM-profile) while keeping the MSA intact.

When such an alignment is available the translation of MSA coordinates into PDB coordinates can be performed easily.

A *Perl* script that automates these steps is available on conditions of GNU General Public License (GPL) at {http://web.mit.edu/~grigory/www/perl/sdr/pdbmsamap.pl}.

Web display. All results, including intermediate ones, are included into the web page report. MSAs are provided as both a FASTA-formatted file and a colored alignment with positions of statistically significant PSDRs highlighted.

PDB structures are displayed using excellent Rasmol/Chime-like Jmol software ({http://jmol.sourceforge.net}). Rasmol script highlighting PSDRs and selecting proper protein models (for NMR-derived structures) and chains is automatically supplied to a Jmol *Java* applet.

Results and Discussion: a case study

We have arbitrarily selected the GerE family of bacterial transcription regulators to illustrate our method. GerE-domains belong to the family of LuxR domains, which, in turn, belong to the larger family of classical 'helix-turn-helix' DNA-binding domains.

We chose the following genomes for our analysis: *Agrobacterium tumefaciens, Aquifex aeolicus, Bradyorhizobium japonicum, Brucella melitensis, Brucella suis, Campylobacter jejuni, Caulobacter crescentus, Escherichia coli, Mesorhizobium loti, Neisseria meningitidis, Rhodopseudomonas palustris, Rickettsia conorii, Rickettsia prowazekii, Sinorhizomium meliloti, Thermoplasma acidophilum, Thermotoga maritima* and *Wolbachia*. The choice of these genomes was not arbitrary but rather dictated by a related study we had been conducting at the time.

We used the GerE PFAM HMM profile (accession number PF00196.8) as an initial input to our program. As the result of the analysis 101 domains were found and clustered into 23 orthologous groups. Eight PSDRs were identified as having significantly high mutual information with p-value < 1.0E-4.

Screenshots of selected pages of the web report are presented in Fig.

As seen on Fig. b seven out of eight PSDRs are indeed directly contacting the DNA. Six of these residues fit tightly into the major groove of the DNA, one contacts the DNA backbone and one is located higher on the surface of the protein. The seven residues contacting a DNA are likely to be responsible for specific DNA site recognition while the one located higher on the surface of the protein can be responsible for contact with other proteins involved in the gene regulation network or can be a false positive.

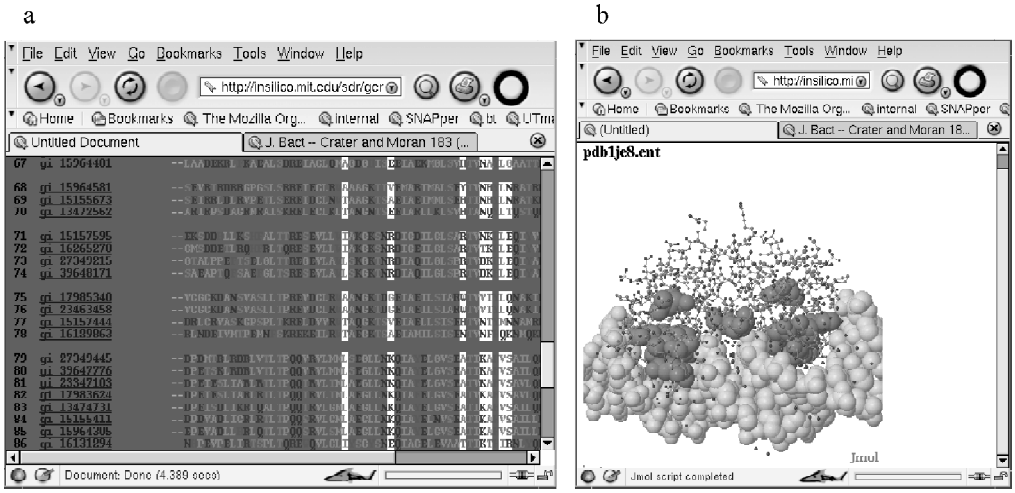


Fig. Screenshots of selected pages of the web report for GerE (grayscaled). a) Multiple sequence alignment, PSDRs positions highlighted. b) 3D structure of GerE-dimer–DNA complex (PDB entry 1je8). DNA (light gray) and PSDRs (dark gray) are rendered in spacefilled mode; the rest of the protein is shown in balls-and-sticks mode.

References

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. 1997. V. 25. P. 3389–3402.

Bateman A., Birney E., Durbin R., Eddy S.R., Howe K.L., Sonnhammer E.L. The Pfam protein families database // Nucleic Acids Res. 2000. V. 28. P. 263–266.

Cover T., Thomas J. Elements of Information Theory, Wiley, New York. 1991.

Eddy S.R. Profile hidden Markov models // Bionformatics. 1998. V. 14. P. 755–763.

Kalinina O.V., Mironov A.A., Gelfand M.S., Rakhmaninova A.B. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families // Protein Sci. 2004. V. 13(2). P. 443–456.

Lichtarge O., Sowa M.E. Evolutionary predictions of binding surfaces and interactions // Curr Opin Struct Biol. 2002. V. 12. P. 21–27

Mirny L.A., Gelfand M.S. Using orthologous and paralogous proteins to identify specificity-determining residues in bacterial transcription factors // J. Mol. Biol. 2002. V. 321. P. 7–20.

Remm M., Storm C.E., Sonnhammer E.L. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons // J. Mol. Biol. 2001. V. 314(5). P. 1041–52.

Thompson J.D., Higgins D.G., Gibson T.J. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. V. 22. P. 4673–4680.

Wollenberg K.R., Atchley W.R. // Proc. Natl Acad. Sci. USA. 2000. V. 97. P. 3288–3291.

MOLECULAR MODELING OF THE NUCLEOTIDE-BINDING DOMAIN OF THE WILSON' DISEASE PROTEIN: THE ATP-BINDING SITE AND DOMAIN DYNAMICS

Kosinsky Yu.A.*[1], Nolde D.E.[1], Tsivkovskii R.[2], Arseniev A.S.[1], Lutsenko S., Efremov R.G.[1]

[1] M.M. Shemyakin & Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Ul. Miklukho-Maklaya, 16/10, Moscow V-437, 117997 GSP, Russia; [2] Department of Biochemistry and Molecular Biology, Oregon Health & Science University, Portland OR, 97239, USA
***** Corresponding author: e-mail: ykos@nmr.ru

Keywords: *copper, ATP7B, P-type ATPase, ATP-binding domain, molecular dynamics, homology modeling*

Summary

Motivation: Wilson's disease protein (WNDP) is a copper-transporting ATPase that plays an essential role in human physiology. Mutations in WNDP result in copper accumulation in tissues and severe hepato-neurological disorder, Wilson's disease. Several mutations were proposed to affect the nucleotide binding and hydrolysis by WNDP, however how the nucleotides bind to normal and mutated WNDP remains unknown.

Results: To aid such studies, we carried out molecular modeling of spatial structure and dynamics of the ATP-binding domain of WNDP and its interactions with ATP. The 3D models of this domain were built using the X-ray structure of the Ca^{2+}-ATPase in the E1 state. To study the functional aspects of the models, they were subjected to long-term molecular dynamics simulations in explicit solvent; similar calculations were performed for the ATP- binding domain of Ca^{2+}-ATPase. In both cases, we found large-scale motions that lead to significant changes of distances between several functionally important residues. The ATP docking revealed two possible modes of ATP binding – *via* adenosine buried in the cleft near residues E1064, H1069, R1151, D1164, and *via* phosphate moiety "anchored" by H-bonds with residues in the vicinity of catalytic D1027. Furthermore, interaction of ATP with both sites occurs if they are spatially close to each other. This may be realized upon relative domain motions of the "closure" type observed in MD simulations. The results provide a framework for analysis of disease mutations and for future mutagenesis studies.

Introduction

WNDP belongs to the large family of the P-type ATPases. During their catalytic cycle, the P-type ATPases became transiently phosphorylated at the invariant Asp residue in the sequence motif DKTG located in the ATP-binding domain (ATP-BD) of the protein. Recently, the high-resolution structure of SR Ca^{2+}-ATPase has been determined in two conformational states [1], providing structural basis for understanding the mechanism of the ATP-driven ion transport.

In contrast to Ca^{2+}-ATPase, very little is known about structure and mechanistic properties of WNDP. Our recent studies demonstrated that mutation H1069Q (the most frequent mutation in Wilson's disease patients) disrupts catalytic phosphorylation from ATP, suggesting that H1069 plays an important role in ATP coordination [2]. However, it remains unclear whether H1069 is the residue coordinating the nucleotide or whether its effect on the ATP binding is indirect. Similarly, several disease-causing mutations were identified in the ATP-BD of WNDP. To better understand the structure and function of the ATP-BD of WNDP, we employed a computational approach, which combines homology modeling, molecular dynamics (MD), and docking simulations.

Methods and Algorithms

The spatial model of the ATP-BD part of WNDP (residues M996-R1322) was built *via* homology modeling. The X-ray structure of the nucleotide-binding domain A320-K758 of Ca^{2+}-ATPase in the E1 ("open") state (PDB entry 1EUL [1]) was taken as a structural template. Construction of the model was done using the Modeller software [3]. The model was subjected to 5-ns molecular dynamics (MD) simulations in explicit water using the GROMACS program [4]. Changes in mutual disposition of the domains were characterized by an angle (θ) between the vectors connecting the center of mass of the hinge region with the centers of mass of the N- and the P-domains. The quality of starting and MD-derived 3D models of WNDP ATP-BD was evaluated by calculation of their 3D_1D compatibility score using the Profiles_3D program [5].

Docking simulations of ATP molecule with the WNDP ATP-BD models were carried out using the GOLD program [6], version 2.0. This was done for the set of 100 conformers extracted with the interval 0.04 ns from the equilibrium parts (1–5 ns) of MD trajectories.

Implementation and Results

It was shown that the overall quality of the constructed 3D models is good enough to employ them for probing the nucleotide binding site and access the role of point mutations in protein-ligand interactions. Strong support for such a conclusion was provided by comparison of computational results with CD data, by inspection of the models with a help of Profiles_3D method, and by MD simulations. The generated model of the WNDP ATP-BD is formed by two sub-domains, the P-domain and the N-domain. The domains are connected by a short linker formed by residues 1033–1038 and 1194–1198. There is a visible cleft in the N-domain, which is adjacent to the linker region. As we show below, this cleft contains a putative ATP-binding site. Accessible surface of the nucleotide-binding site is formed by non-polar side chains, except that of D1164.

Two selected models of the WNDP ATP-BD were subjected to 6-ns MD simulations in explicit water solution. These simulations pursued the following goals: 1) To inspect the global and the local conformational lability of the models, especially, near the loop 1062–1071; 2) To generate a set of equilibrated MD conformers for subsequent docking with ATP and to determine how conformational dynamics affects the ability of models to bind the ligand. The MD simulations revealed that the conclusions made for Ca^{2+}-ATPase and WNDP ATP-BD quite similar. Thus, the N- and P- domains in WNDP do not unfold during MD and retain their spatial structure well – the corresponding root-mean-square deviations between coordinates of the backbone atoms of the MD- and starting structures are 3.0–3.5 Å . Analysis of the equilibrium angles θ obtained in both trajectories shows that the protein models built in the "open" form converge to rather more "closed" conformations: θ reaches ~115°–120° from the starting value of 160°. In addition, this global conformational transition leads to increasing of the inter-domain surface area from ~30 Å to ~100 $Å^2$. The domains interact *via* the loop 1152-1156 in the N-domain and α-helix 1225–1232 in the P-domain. It seems that such a contact may stabilize a certain orientation of functionally important residues in WNDP.

Exploration of the conformational space *via* MD simulations provides an opportunity to examine binding of ATP by protein present in different equilibrium states. The results of docking to the set of MD-conformers may be divided into two groups:

1. Adenosine part binds in the cleft of the N domain Majority of the complexes possess H-bond between the NH_2 group of ATP and carboxyl group of D1164. Side chains of E1064 and H1069 (mutations of these residues lead to the Wilson's disease) have tight contacts with the ribose moiety. Moreover, the first of them creates H-bond with the O2'H-group of ribose. Also, a transient

H-bond between side chains of these two residues was observed in MD runs. Probably, it stabilizes the overall conformation of the loop and its orientation with respect to the binding site. The phosphate tail of ATP forms one or two H-bonds with the side chain of R1151 – this residue approaches the binding site in the result of domain motions observed in MD. Together with residues K1028 and D1222 from the P domain, it "anchors" the phosphate moiety of ATP and orients it towards functionally important D1027. On the other hand, γ-phosphate of the ligand is still far (> 8 Å) from the phosphorylation site.

2. The γ-PO_3 group of ATP is in direct contact with D1027, while the adenosine moiety binds on the interface between the N and P domains. In this case, the protein-ligand interaction is determined by the network of H-bonds between the phosphate groups of ATP and residues D1027, K1028, T1029, T1220 of the P domain.

To summarize, WNDP ATP-BD contains two distinct motifs responsible for ATP binding. These include the hydrophobic cleft in the N-domain that shields the adenosine moiety from solvent, and the cluster of residues in the P-domain, which "anchor" the phosphate tail of ATP *via* the network of H-bonds. The high-affinity binding of ATP is likely to occur upon simultaneous contacts with both motifs. Spatial rapprochement of the binding sites takes place as the result of the domains "closure" observed in MD. A number of residues potentially involved in coordination of ATP were delineated. Their predicted role will be tested in the future site-directed mutagenesis studies. Overall agreement of the simulation results with the experimental data makes the modeling a powerful tool that may be applied to explore spatial structure, dynamics, and functioning of the ATP-BD part of WNDP, its mutants, and homologous enzymes.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (Grant 04-04-48875-a), by the Ministry of Science and Technology of the Russian Federation (the State contract 43.073.1.1.1508/31.01.2002), and by the National Institute of Health Grant PPG 1-P01-GM067166-01 to S.L. R.G.E. and D.E.N. are grateful to the Science Support Foundation (Russia) for the grants awarded.

References

1. Toyoshima C., Nakasako M., Nomura H., Ogawa H. Crystal structure of the calcium pump of sarcoplasmic reticulum at 2.6 A resolution // Nature. 2000. V. 405. P. 647–655.
2. Tsivkovskii R., Efremov R.G., Lutsenko S. The role of invariant His1069 in folding and function of the Wilson's disease protein, the human copper-transporting ATPase ATP7B // J. Biol. Chem. 2003. V. 278. P. 13302–13308.
3. Sali A., Overington J.P. Derivation of rules for comparative protein modeling from a database of protein structure alignments // Protein Sci. 1994. V. 3. P. 1582–1596.
4. Berendsen H.J.C., van der Spoel D., van Drunen R. GROMACS // Comp. Phys. Comm. 1995. V. 91. P. 43–56.
5. Lüthy R., Bowie J.U., Eisenberg D. Assessment of protein models with three-dimensional profiles // Nature. 1992. V. 356. P. 83–85.
6. Jones G., Willett P., Glen R.C., Leach A.R., Taylor R. Development and validation of a genetic algorithm for flexible docking // J. Mol. Biol. 1997. V. 267. P. 727–748.

ANALYSIS OF PROTEOME COMPEXITY BASED ON COUNTING DOMAIN-TO-PROTEIN LINKS

Kuznetsov V.A.*[1], Pickalov V.V.[2], Knott G.D.[3], Kanapin A.A.[4]

[1] CIT/NIH & SRA International, Inc. Bethesda, MD, 20892, USA, e-mail: vk28u@nih.gov; [2] Institute of Theoretical and Applied Mechanics SB RAS, Novosibirsk, Russia, e-mail: pickalov@itam.nsc.ru; [3] Civilized Software, Inc., Silver Spring, MD, 20906, USA; [4] EMBL-EBI Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK,
* Corresponding author: e-mail: alex@ebi.ac.uk

Keywords: *proteome complexity, evolution, domain-to-protein links, skew distributions*

Summary

We define the list of domains to proteins, together with the numbers of their occurrences (links to proteins) found in the proteome of an organism to be the *domain-to-protein linkage profile* (DPLP) of the proteome. We estimated the DPLP for the 156 fully-sequenced genome organisms represented in the InterPro database. This work presents several quantitative measures of the complexity of a proteome based on the DPLP. For each of the 156 studied organisms, we found two large subsets of domains: the domains which occur two or more times in at least one protein, and the domains which are not duplicated within any protein of the proteome. The latter set of domains well reflects the increasing trend of biological complexity due to evolution. The statistical distributions of the number of domain-to-protein links in the proteome and the estimates of the differences between the DPLPs for pairs of organisms are used as measures of relative biological complexity of the organisms. In particular, we show quantitatively the greater complexity of the human proteome, relative to that of a mouse or a rat. These differences are only partially reflected in the number of protein-coding genes estimated for these species by the sequencing genome projects.

Introduction

There are no consensus quantitative measures of the relative or absolute complexity of an organism. The biological complexity of an organism should be characterized by the list of all proteins in a given proteome and their dynamic interactions among themselves (protein network), the other molecules (protein-DNA network, etc.). However, the total number of putative protein sequences based on complete genome sequence data of a given organism can be predicted only approximately, and our knowledge on dynamic interactions of proteins in an organism is also essentially incomplete. Thus, we need a more tractable and practical approaches to describing the biological complexity.

Proteins contain short structural and/or functional 'blocks' (sequences of amino acids) called structural motifs (of length 5–35 nt) and structural domains (of length 30–250 nt) that are seen repeatedly in many proteins of all species whose genome have been examined. The entire number of domains in nature is probably a very limited number. The estimated number of such classes of homologous sequences of motifs and domains in nature ranges from 6,000 to 10,000 or so [1–3]. In this work, all such protein blocks will be called, for brevity, *domains*. The domains are essential to the biological function (s) of the protein in which they occur and serve as evolutionarily conserved building blocks for the forming of proteins. The domains are corresponding to specific sequences of DNA within genes which have been evolutionarily conserved. DNA corresponding to a domain may occur multiple times in a given protein-coding gene and/or in many different protein-coding genes within a given genome, and in the genomes of many species. We attempt to determine the measures of the biological complexity based on a limited set of domains.

If a domain D_i occurs in the protein P_j, we say this constitutes a *domain-to-protein link*. The list of domains together with the numbers of occurrences of each domain (links) found in the proteome of an organism is called the *domain-to-protein linkage profile* (DPLP) of the proteome. The DPLP should allow us to characterize the domain-to-protein link network for each organism. Currently, we do not know all domains and all proteins in a given organism. We can study the DPLP of organism by observing sample DPLPs in representative proteomes (i.e. protein sets which were specified for organisms). However, several protein domain/motif databases provide large enough samples of DPLPs for a variety of organisms. We can analyze the samples of domain-to-protein linkage information in fully-sequenced genome organisms to categorize their proteome complexities. This work presents several quantitative measures of the complexity of a proteome based on observed DPLPs that appear to be appropriate and consistent.

Protein Domain Database Analyzer

Our information about DPLPs of the 156 fully-sequenced genome archaeal, bacterial and eukaryotic organisms was obtained from the InterPro database (www.ebi.ac.uk/interpro, April, 2004) [5], the protein sequences for the organisms were obtained from the Proteome Analysis Database (www.ebi.ac.uk/proteome, April 2004).
We developed a Protein Domain Database Analyzer (PDDA) program which we used to access the InterPro database and download data into a local MySQL relational database [1]. The MySQL database consists of a $N_{tot} \times L$ table, where N_{tot} = 9609 domains; and L = 156 organisms. Each row corresponds to an InterPro domain and each column corresponds to an organism. The (i, S) -th entry of the table is the number of occurrences of the domain i in the proteome of organism S. Information of this table was analyzed by PDDA data mining tools, which include the statistical and graphical functions, logical functions, descriptive statistics, and correlation analysis.

Definitions

Let us focus on a specific organism S. Let $A = \{a_1, a_2, ...a_N\}$ be the set of observed domains contained in the P proteins of the organism S. Let $B = \{b_1, b_2, ..., b_P\}$ be the set of proteins in the representative proteome (sample of observed proteins) for the organism S. We define the $N \times P$ adjacency matrix $C' := [c'_{ij}]$ ($i = 1,2,..., N$; $j = 1,2,..., P$) for the organism S as follows: the value $c'_{ij} = k$, if protein b_j contains domain a_i exactly k times. Note that $k \in \{0,1,2,...\}$. The value c'_{ij} is the number of domain-to-protein links for the (domain, protein) pair $\{a_i, b_j\}$.

Let $m'_i := C'_{i*} = \sum_{j=1}^{P} c'_{ij}$ denote the number of occurrences of the domain a_i in all proteins of the representative proteome of S; $i = 1,2,..., N$. We call the value m_i the number of r*edundant domain-to-protein links* of the domain a_i of S. Let M' denote the total number of domain-to-protein links in the representative proteome of the organism S. $M' = \sum_{i=1}^{N} \sum_{j=1}^{P} c'_{ij}$. The value M' is

called the redundant connectivity number of the (domain, protein) network. Note that all occurrences of a domain in the proteins of *S* are counted.

We also define the adjacency matrix *C*, where $c_{ij} = 1$, if protein b_j contains domain a_i at least once, and we define $c_{ij} = 0$ otherwise. This matrix reflects the non-redundant structure of the domain-to-protein links in the (domain, protein) network. This matrix counts each (domain, protein) link only once. Let $m_i = \sum_{j=1}^{P} c_{ij}$. We call the value m_i the number of *non-redundant domain-to-protein links* of the domain a_i of *S*. Let $M = \sum_{i=1}^{N}\sum_{j=1}^{P} c_{ij}$. Note $M \le M'$. We call the matrix *C* the non-redundant adjacency matrix of *S*, and we call *M* the non-redundant connectivity number of the (domain, protein) network of *S*. The values *M* and *M'* have been used as the non redundant and redundant measures of the proteome complexity of an organism [1, 4].

To quantify the complexity associated with multiple occurrences of the domain a_i in an organism, we define the redundancy value

$$\delta m_i = m'_i - m_i,$$

(1) To quantify the relative complexity of the domain a_i which redundantly occurs $m_i{}'^{(s)}$ times in the organism S and which redundantly occurs $m'^{(k)}{}_i$ times in the organism *K*, we can define the difference of redundant complexity value for the domain a_i

$$\delta m'_i{}^{(s,k)} = m'_i{}^{(s)} - m'_i{}^{(k)}.$$

By omitting the prim symbol in above notations, we can also introduce the difference of non redundant complexity value of the domain a_i

$$\delta m_i^{(s,k)} = m_i^{(s)} - m_i^{(k)}.$$

To characterize the relative redundant complexity of the organism S versus the organism *K*, we use the empirical bivariate distributions of the random values $(\mu'^{(s,k)}_i, \delta m'^{(s,k)}_i)$, where $\mu'^{(s,k)}_i = (m'^{(s,k)}_i + m'^{(s,k)}_i)/2$ is a mean value of the numbers of the links and $i = 1,2,...,N^{(s,k)}$; $N^{(s,k)}$ is the total number of domains occurred in domain-to-protein profiles of the organisms *S* and *K*. By omitting the prim symbol in above notations, we obtain the empirical bivariate distribution of the random values $(\mu_i^{(s,k)}, \delta m_i^{(s,k)})$, which we can use as the relative non redundant complexity measure of the organisms *S* and *K*.

Results and Discussion

There are at least two distinguish processes of domain's spread in nature: integration of two or more different domains in a new protein (forming non-redundant domain-to-protein links) and multiplication of a domain within the same protein (forming redundant domain-to-protein links). Panels A-C in Figures 1 show the relationships between the numbers of non-redundant and the numbers of redundant domain-to-protein links counted for the *T. volcanium*, *E. coli* and human

representative proteomes. These panels demonstrate typical patterns of the bivariate frequency distributions of the random values (m,m') $(=\{(m_1,m'_1),...,(m_{N^{(s)}},m'_{N^{(s)}})\})$ corresponding to the numbers of redundant and non-redundant occurrences of domains in the organism S. All panels in Figure 1 show a preferential distribution of points along the diagonal; the spread of points in the orthogonal direction to the diagonal is smaller. The extreme cases (*T. volcanium* and human proteomes) clearly display the increased complexity of the human proteome due to the increased number of proteins which preferentially combining different domains together. Panel D in Fig. 1 shows the observed values (m,m') for the 156 organisms is a factor of 10 larger then any specific organism. We found that these distribution patterns are typical for the studied archael, bacterial and eukaryotic organisms. Figures 1A-D imply that increasing biological complexity in nature is mostly associated with formation of new multi-domain proteins combining different domains rater that with the increase of domain-to-protein links occurring due to increasing of multiplication of domains within proteins in which domains have been already occurred.

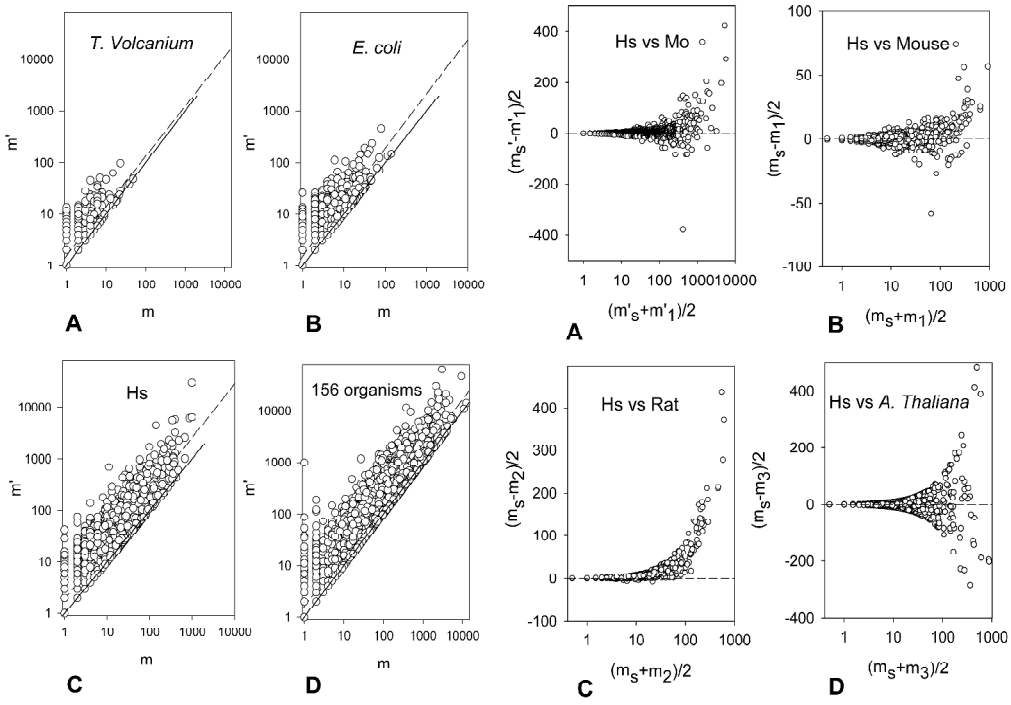


Fig. 1. The empirical distributions of the numbers of redundant links versus the numbers of non redundant links counted for the (A) *T. volcanium*, (B) *E. coli* and (C) human representative proteomes, and for (D) pooled data of 156 organisms. Discontinue line: linear regression; solid line: diagonal.

Fig. 2. The relative complexity of proteome of a human versus a mouse, a rat (C) and *A. thalina*, respectively. A: the distribution of $(\mu'^{(s,k)}_i, \delta m'^{(s,k)}_i)$; symbol *s* indicates the human organism, k = 1 indicates the mouse organism. B, C, D: the distributions of $(\mu_i^{(s,k)}, \delta m_i^{(s,k)})$; k=2, 3, 4 indicate the mouse, the rat and *A. thalina*, respectively. $i = 1,2,..., N^{(s,k)}$.

For all of the 156 studied organisms, we also found two disjointed subsets of domains: the domains which occurred two or more times in at least one protein of a proteome, and the domains which was not multiplied within any protein of the proteome (see examples on Fig. 1). 4541 (47 %) of the 9609 domains occurred two or more times in a same protein of the 156 organisms. We also found that if a domain is more common in the entire proteome world, then that domain has a bigger chance to appear multiple times in a protein of any organism (see Fig. 1).

Figure 2 demonstrates how the proteome complexity of pairs of organisms can be compared. This figure shows the relative proteome complexity measure for a human versus a mouse, a rat and *A. thaliana*, respectively. For example, the panel A shows the bivariate distributions of differences of the number of redundant links for the human and mouse organisms, multiplied by factor 0.5 with respect to average number of the redundant links of the human and mouse organisms. This distribution is highly asymmetric about axes x: the number of positive differences (abundant for a human) is bigger than the number of negative differences (abundant for a mouse), and the positive highly-abundant values occur more often in the human sample than in the mouse sample. Similar asymmetric trend we observed for our human-mouse comparison in the case of redundant links (Fig. 2B). This indicates that the human organism reuses domains more frequently in same protein and in different proteins, and invents more diverse multi-domain proteins than the mouse organism. This analysis suggests that even though that the numbers of non redundant protein-coding genes in a human (~33,600 genes [1, 4]) and in a mouse (~32,000 genes, by our current estimate based on the method in [1, 4]) are approximately similar, and that even the total number of domains are also similar numbers in these organisms (~5,600 for a human and ~5,100 for a mouse [4]), the proteome complexity and diversity of a significant fraction of multi-domain proteins in a human is higher than in a mouse.

Large asymmetry in the bivariate distribution of the values ($\mu_i^{(s,k)}$, $\delta m_i^{(s,k)}$) around axes x we observed for a human versus a rat (Fig. 2C). However, the human and *A. thaliana* proteomes show similar proteome complexity by our criteria (Fig. 2D). This is not a surprise, because even though the number of protein-coding genes in human is larger than the number of protein-coding genes in *A. thaliana* (~26,000-27,000 [1, 4]), relatively recent massive (and imperfect) genome duplication events in *A. thaliana* might dramatically increased of protein repertoire due to recombination events, which perhaps increased of the domain shuffling in proteins, but did not increase of the number of domains (~5,300 domains [4]) presented in the *A. thaliana* proteome.

References

1. Kuznetsov V.A. Statistics of the numbers of transcripts and protein sequences encoded in the genome // Computational and Statistical Methods to Genomics / Eds. Zhang W., Shmulevich I. Kluwer, Boston, 2002. P. 125–171.
2. Kuznetsov V.A., Pickalov V.V., Senko O.V., Knott G.D. Analysis of evolving proteomes: Predictions of the number of protein domains in nature and the number of genes in eukaryotic organisms // J. Biol. Systems. 2002. V. 10. P. 381–407.
3. Kuznetsov V.A. Family of skewed distributions associated with the gene expression and proteome evolution // Signal Processing. 2003a. V. 83. P. 889–910.
4. Kuznetsov V.A. A stochastic model of evolution of conserved protein coding sequence in the archaeal, bacterial and eukaryotic proteomes // Fluctuation and Noise Letters. 2003b. V. 3. L295–L324.
5. Mulder N.J., Apweiler R., Attwood T.K. *et al*. InterPro Consortium. The InterPro Database, 2003 brings increased coverage and new features // Nucl. Acids Res. 2003. V. 31. P. 315–318.

VALIDATION OF RANDOM BIRTH-DEATH MODEL OF EVOLUTION OF PROTEOME COMPLEXITY

Kuznetsov V.A.

CIT/NIH & SRA International, Inc. Bethesda, MD, 20892, USA, e-mail: vk28u@nih.gov

Summary

We have shown that counting the domain-to-protein links observed in the protein and protein domain/motifs data sets and analysis of statistical distributions of these counts lead to common probabilistic model of evolution of proteome complexity of the archael, bacterial and eukaryotic organisms (Kuznetsov *et al*., 2002; Kuznetsov, 2003a, b). In this work, using InterPro data sets, we test several basic assumptions and predictions of our model.

Keywords: *proteome complexity, evolution, domain-to-protein links, skew distributions*

Introduction

Proteins contain short structural and/or functional 'blocks' (sequences of amino acids) called structural motifs (of length 5–35 nt) and structural domains (of length 30–250 nt) that are seen repeatedly in many proteins of all species whose genome have been examined (Kuznetsov *et al*., 2002; Kuznetsov, 2003a). In this work, all such protein blocks will be called, for brevity, *domains*. The domains are correspond to specific sequences of DNA within genes which have been evolutionarily conserved. DNA corresponding to a domain may occur multiple times in a given protein-coding gene and/or in many different protein-coding genes within a given genome, and in the genomes of many species.

If a domain a_i occurs in the protein b_j, we say this constitutes a *domain-to-protein link*. The list of domains together with the numbers of occurrences of each domain (links) found in the proteome of an organism is called the *domain-to-protein linkage profile* (DPLP) of the proteome. The DPLPs allow us to characterize the domain-to-protein link network (and proteome complexity) for each organism *S*.

Let us focus on a specific organism *S*. Let the random variable X' denote the number of occurrences of the domain a_i in all proteins of the representative proteome of *S*. We call this number as the number of *redundant* domain-to-protein links of a random domain of the organism *S*. We can define the domain occurrence probability function (DOPF) $p'_l := \Pr(X' = l)$, where $l = 0,1,2,\ldots$.

The value p_l is the probability that a random domain occurs exactly *l* times within the proteome of the organism.

If protein b_j contains domain a_i at least once, we can count the number of such non-redundant occurrences (non-redundant domain-to-protein links) in the (domain, protein) network for the given domain. Let the random variable *X* denote the number of *non-redundant* domain-to-protein links of a random domain of *S*. We also define the DOPF $p_h := \Pr(X = h)$, where $h = 0,1,2,\ldots$. The value p_h is the probability that a random domain occurs exactly *h* times within the proteome of *S*.

Let $A = \{a_1, a_2, \ldots a_N\}$ be the set of observed domains contained in the P proteins of the organism *S*.

Let $B = \{b_1, b_2, \ldots, b_P\}$ be the set of proteins in the representative proteome (sample of observed

proteins) for the organism *S*. The value $\hat{p}'_l = \hat{n}'_l / N$ $(l = 1,2,...)$ is an empirical estimate of the probability function p'_l, where $\hat{n}_l$ is the number of observed domains occurring exactly *l* times in the representative proteome of *S* and *N* is the total number of observed domains in the representative proteome. Let M' denote the total number of domain-to-protein links in the representative proteome of the organism *S*. We call the value M' the redundant connectivity number of the (domain, protein) network. Note that all occurrences of a domain in the proteins of *S* are counted.

As before we can obtain an estimate of p_h. The value $\hat{p}_h = \hat{n}_h / N$ $(h = 1,2,...)$ is an empirical estimate of the probability function p_h, where $\hat{n}_h$ is the number of non redundant domain-to-protein links occurring exactly *h* times in the representative proteome of S. Let *M* the total number of non redundant domain-to-protein links in the representative proteome of the organism *S*, and we call *M* the non-redundant connectivity number of the (domain, protein) network of *S*. Note $M \le M'$. The values *M* and M' can be used as the non redundant and redundant measures of the proteome complexity of an organism (Kuznetsov *et al*., 2002; Kuznetsov, 2003a).

Let $X^{(s)}$ be the number of non redundant links for a random domain of the organism *S*. Let $X^{(k)}$ be the number of non redundant links for a random domain of the other organisms. To characterize the relative non-redundant complexity of the organism *S* versus the organism *K*, we can plot the bivariate distributions of the random variables $(X^{(s)}, X^{(k)})$.

In this work, we test the assumptions and predictions of our probabilistic model of evolution of proteome complexity based on the analysis of statistical distributions of the redundant and non redundant domain-to-protein link networks of variety organisms. For this purpose, data sets of the 156 fully-sequenced genome organisms was downloaded from the InterPro database (www.ebi.ac.uk/interpro) and from the Proteome Analysis Database (www.ebi.ac.uk/proteome).

Hypergeometric Model of Evolution

Let the random variable $D_t(a, P)$ be the number of domain-to-protein links of the domain a in the organism occurring at time *t* in the evolutionary path *P* of some end-point organism. The $D_t(a, P)$ is a realization of a continuous-time stochastic process $\{D_t, t \ge 0\}$. This process can be considered as a birth-death random process that has a net change across an infinitesimal time interval. Note for redundant links $\sum_{i=1}^{N} m'_i = M'$ and for non redundant links $\sum_{i=1}^{N} m_i = M$. Let J denote the number of non redundant links for most abundant domains found in the proteome and let J' denote the number of occurrences of the most abundant domains found in the proteome. Note $J' \ge J$ and the both values *J* and J' are increasing if *M* becomes bigger.

Let consider the case of non-redundant domain-to-protein links. Let $p_m(t) = P(D_t = m)$ $(m = 0,1,2,...,J)$ denote the probability function associated with the random process $\{D_t, t \ge 0\}$. Then the rate of the probability function p_m $(m = 0,1,2,...,J)$ can be described by finite system of the forward Kolmogorov equations. For our application purposes, we consider the intensity rates of the random process $\{D_t, t \ge 0\}$ be the linear functions of the value *m*:

$\lambda_s = \lambda_1^* + \lambda_2^* m$

and

$\mu_s = \mu_1^* + \mu_2^* m$,

where $\lambda_1^* > 0, \lambda_2^* > 0, \mu_1^* > 0, \mu_2^* > 0$. Hence, during an interval $(t, t+\tau)$ where τ is small, we assume four independent processes: the spontaneous "birth" and "death" of a domain, with constant intensities λ_1^* and μ_1^*, respectively, and the "flows" of the domains with the intensities proportional to the number of the links already counted for that domain $\lambda_2^* m$ and $\mu_2^* m$. The parameters λ_2^* might be associated with purification and positive selection of already used links for a random domain and μ_2^* might be associated with loss of the links due to random constraints and negative selection forces.

We assumed that in the most evolving near end-point organisms, the random birth and death processes of protein-encoding sequences are keeping near equilibrium. This equilibrium can be parametrically described with the following recursive probability formula which we called the Kolmogorov-Waring (KW) probabily function (Kuznetsov, 2003a, b).

$$p^*_{m+1} = \theta \frac{(a+m)}{b+m+1} p_m^* , \qquad (1)$$

where $m = 0,1,...$, and a, b, and θ are the positive parameters of our model;

$p_0 = \frac{1}{{}_2F_1(a,1;b+1;\theta)} > 0$, where ${}_2F_1(a,1;b+1;\theta)$ is the hypergeometric Gauss function.

$a = \lambda_1^* / \lambda_2^*$, $\theta = \lambda_2^* / \mu_2^*, b = \mu_1^* / \mu_2^*$.

In order to fitting the probability function Eq.1 to the empirical DOPFs, we re-normalized this probability function due to the random values of empirical DOPFs are changed between 1 to J and estimated of the parameters a, b, θ, as described in (Kuznetsov, 2003 a, b).

Results

Even with their large differences in proteome complexity of archael, bacterial and eukaryotic organisms all demonstrate similar skewed long-tail (Pareto-like) DOPFs (see three examples on Fig. 1A) and all data sets fit accurately by Eq.1. The shape of the empirical DOPFs, the parameter J and estimated parameters of the best-fit probability function by Eq. 1 correlate on the total number of domain-to-protein links M. Note that for most of the studied organisms (with several exceptions among bacteria and archaea) estimated values of the parameter θ equal around 1, which means that $\lambda_2^* \approx \mu_2^*$. This results specifies our assumption that rates of the birth and death processes should be near equilibrium. However, the best-fit model predicts that for many bacterial and arcaheal organisms the equilibrium in domain-to-protein link network might be unstable and the dispersion approaches infinity.

To characterize the relative redundant complexity of the organisms S versus the organism K, we used the empirical bivariate distributions of the random variables $(X^{(s)} = m^{(s)}, X^{(k)} = m^{(k)})$, where $m^{(s)} = 0,1,2,..., J^{(s)}$ $m^{(k)} = 0,1,..., J^{(k)}$. Figure 1B shows typical pattern of such distribution when more complex organism (*Drosophila melanogaster*) compares to the less complex organism (yeast): a random domain which occurs in more complex (multi-domain) proteins of

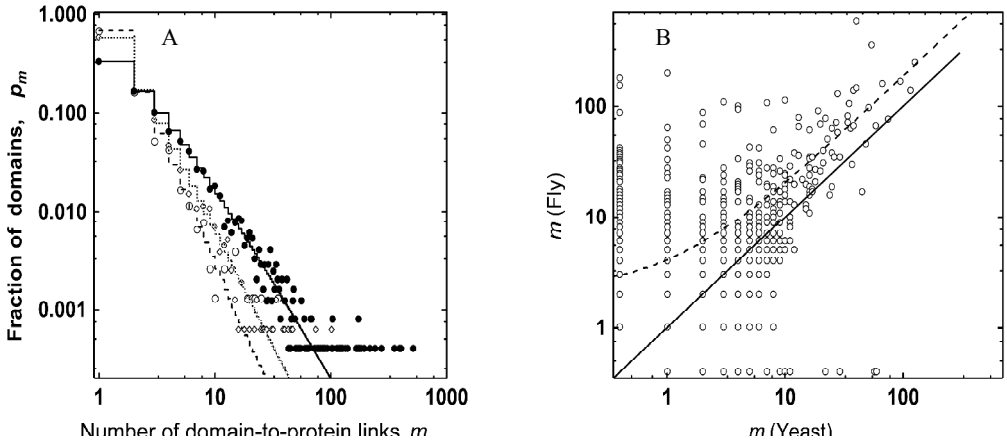


Fig. 1. Analysis of the empirical distributions. A: Fitting of Eq.1 to the empirical DOPFs for the archael, bacterial, and eukaryotic representative proteomes. o,◊,•: *T. volcanium*, *S. typhi*, respectively; m=1,2,...,*J. Step-functions*: best-fit Eq.1 at a=0.017; b=2.412; θ=1.046 for data o; best-fit Eq.1 at a=0.63; b=1.198; θ=1.00 for data ◊; best-fit Eq.1 at a=0.433; b=0.653 and θ=0.98 for data • (see also Methods in [2,3]). B: bivariate distribution of the number of proteins containing the same domain in the yeast and *D. melanogaster* proteomes. *Solid line*: diagonal; *discontinue line*: regression line.

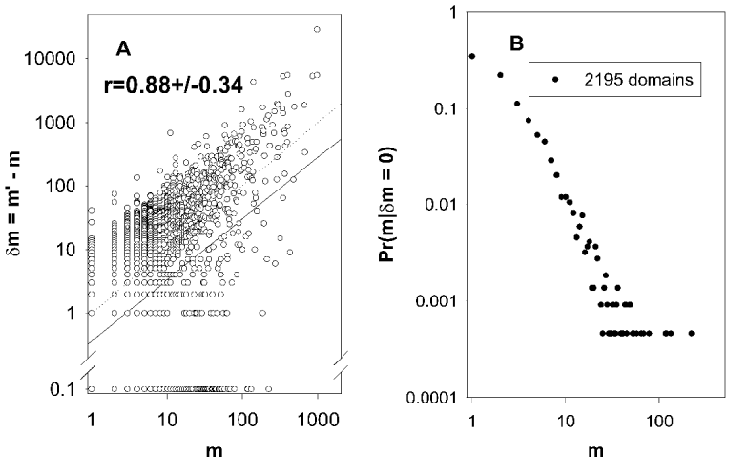


Fig. 2. Analysis of complexity of human proteiome. A: relationships of the differences between the number of redundant domain-to-protein links and the number of non-redundant domain-to-protein links in human representative proteome. B: empirical conditional probability distribution of no multiple occurrences of the domains in the same protein among domains non-redundantly occurred one, two, and more times in the human representative proteome.

less complex organism has a higher than average chance to appear in many more complex proteins in a more complex organism, even over widely evolutionary paths. Figure 1B shows that domains represented by a larger number of non-redundant links in the yeast proteome tend to also more often in the fly proteome. We found that 'yeast' domains (evolutionarily "older" domains) repre-

sented in the fly proteome have, on an average, more domain-to-protein links than domains represented only in the fly proteome (domains not presented in the yeast proteome; perhaps, evolutionarily "younger" domains). These results indicate that the probability of acquisition of new non-redundant links to proteins is roughly proportional to the number of occurrences of this domain (in particular in the "younger" proteomes). This conclusion qualitatively agrees to the basic assumptions lead to Eq.1.

Figure 2A shows strong positive correlation ($r=0.88$; $p<0.001$) between the random increment of domain-to-protein links for the domains of the organism ($\delta m = m' - m$) and the number of non redundant domain-to-protein links (m) in that organism. This data supports the analysis presented in Figure 1B.

Figure 2B shows the empirical conditional probability of random variable $(m \mid \delta m = 0)$ which is the number of links of a random domain among the domains which has no multiple occurrences $(\delta m = 0)$ in any protein. 2195 (47 %) of the 4705 human domains have no multiple concurrencies. The form of histogram in Figure 2B fits well by Eq.1 (not presented), and allows us to assume that redundant occurrence of rare and/or recent domains within proteome of an organism is much less than the redundant occurrence of more abundant domains. This result is qualitatively support our simple probabilistic model of evolution of proteome complexity.

Conclusion

We might assume that the domain occurrence counts in a proteome during evolution is a random birth-death quasi-steady state process such that new domains are rarely appeared as a singletons and lost at constant rates, and domains are reused and lost at a rates proportional to their current use. The occurrences of a given domain in the proteome is essentially random process, determined by the intrinsic properties of domain (i.e. hydrophobic group, etc.) and also the evolution history of the domain (i.e. evolution age). We need to better understand the non-random mechanisms lead to changes of the skewed distribution of domain-to-protein links and to specific domain-domain interactions in the course of evolution.

References

Kuznetsov V.A., Pickalov V.V., Senko O.V., Knott G.D. Analysis of evolving proteomes: Predictions of the number of protein domains in nature and the number of genes in eukaryotic organisms // J. Biol. Systems. 2002. V. 10. P. 381–407.

Kuznetsov V.A. Family of skewed distributions associated with the gene expression and proteome evolution // Signal Processing. 2003a. V. 83. P. 889–910.

Kuznetsov V.A. A stochastic model of evolution of conserved protein coding sequence in the archaeal, bacterial and eukaryotic proteomes // Fluctuation and Noise Letters. 2003b. V. 3(3). L295–L324.

INFORMATION ABOUT SECONDARY STRUCTURE IMPROVES QUALITY OF PROTEIN ALIGNMENT

Litvinov I.I.*[1], Mironov A.A.[2], Finkelstein A.V.[3], Roytberg M.A.*[1]

[1] Institute of Mathematical Problems in Biology RAS, Puschino, Russia; [2] Moscow State University, Department of Biotechnology and Bioinformatics, Moscow, Russia; [3] Institute of Protein Research RAS, Puschino, Russia

***** Corresponding authors: e-mail: litvinov@mail.ru, Roytberg@impb.psn.ru

Keywords: *protein sequence alignment, secondary structure prediction*

Summary

Motivation: The Smith-Waterman (SW) alignment algorithm is known as the most accurate algorithm for pair wise alignment of amino acid sequences.. It means, that SW alignments are more similar to alignments of corresponding 3D-structures, than FASTA, BLAST, etc. alignments. But even SW algorithm is unable to restore alignment of proteins' 3D-structures if the sequence identity is less than 30 % ("twilight zone"). Our goal is to design a new alignment method, which is significantly more accurate, than SW algorithm.

Results: We propose to modify SW alignment score to take into account protein secondary structure. We give bonus for alignment of residues belonging to the regions of same secondary structure type. We have shown that alignments maximizing the improved score are much more accurate, than SW alignments (57 % accuracy vs. 31 % for the twilight zone sequence identity; both experimentally determined and theoretically predicted secondary structure can be used). The dynamic programming algorithm to find the optimal secondary structure alignment was designed and implemented as C++ program STRUSWER.

Availability: Program STRUSWER is available on request from the authors.

Introduction

Alignment of amino acid sequences is the core of many modern bioinformatics methods, e.g. homology based 3D-modeling of proteins, database search, etc. The algorithmically produced protein alignments usually differ from the "structure" alignments, i.e. alignments obtained by superposition of 3D-structures. The latter can be considered as "golden standard" ones because of two reasons. First, protein 3D-structures is much more conservative, than their amino acid sequences, and therefore structure alignments are more close to evolutionary based alignments. Second, for homology based 3D-modeling of proteins and many other applications, it is important to restore 3D-structure alignments.

It is known, that Smith-Waterman (SW) alignment algorithm [6] creates alignments which are most similar to the structure alignments compared to alignments produced by other popular alignment algorithms, e.g. BLAST [1, 2], FASTA [5], etc. But even SW algorithm is unable to restore alignment of proteins' 3D-structures if the sequence identity is less than 30 % [7].

To improve the quality of SW algorithm we propose to take into account proteins' secondary structure. Both experimentally determined and theoretically predicted structures can be used. The predictions can be done both in "strict form", i.e. ascribing each residue with mark H (for helix); E (for beta-strand) and L (loop), and in "propensity form", i.e. ascribing each residue with probabiliities to belong to a helix, strand or loop. The program STRUSWER, which implements the approach have shown up significantly better quality than standard implementation of SW algorithm.

Materials and Methods

Structure alignments. We use manually verified structure alignments from the BAliBase [8] protein structure database, as a source of "golden standard" alignments. BAliBase contains manually curated multiple alignments, initially based on 3D protein structures. We have used alignments from BAliBase Reference 1, the sequence identity level for the Reference is mainly 10–50 %. The test set was consisted of all protein pairs meeting following condition: both proteins belong to the same multiple alignment of BAliBase's Reference 1 and their 3D-structures are known.

Secondary structure. Experimentally determined structure was obtained from database DSSP [4]. The 8-state DSSP description of a structure was converted to 3-state structure form using following rules:

(H,G,I)->H; (E,B)->E; (T,S)->L,

where H = alpha helix; B = residue in isolated beta-bridge; E = extended strand, participates in beta ladder; G = 3-helix (3/10 helix) ; I = 5 helix (pi helix); T = hydrogen bonded turn ; S = bend.

To predict secondary structures we have used the PsiPred [3] program in its full (using homology search technique), and restricted (using only amino acid itself, AKA "PsiPred single") versions. We have used two forms of secondary structure representation. First ("strict form"): each residue R_i is ascribed with a mark $T(R_i)$; the possible values for T_i :are 'H' (for a helix); 'E' (for a beta-strand) and 'L' (for a loop). This form is also used to represent the experimentally determined secondary structures. Second, ("propensity form"): each residue. R_i is ascribed with a probabilities $P(R_i, T)$; that R_i belongs to a secondary structure of type T, i.e. a helix, a strand or a loop. Below we use following abbreviations for the above types of secondary structure assignments: a) Exp – Experimentally determined structure; b) Psi – the structure, predicted by the full version of PsiPred; strict representation of secondary structure is used; c) PsiPro – the structure, predicted by the full version of PsiPred; propensity representation of secondary structure is used; d) PPS – the structure, predicted by the restricted version of PsiPred; strict representation of secondary structure is used; e) PPSPro – the structure, predicted by the restricted version of PsiPred; propensity representation of secondary structure is used.

Alignment score and alignment algorithm. Original Smith-Waterman alignment score [8] was modified as follows. We define score $W[a_i, b_j]$ of matching of residues a_i from the sequence A and b_j from the sequence B is defined as:

$$W(a_i, b_j) = M(a_i, b_j) + SBON*Q(a_i, b_j),$$

where $M(a_i, b_j)$ is a substitution score given by substitution matrix (e.g. blosum62), SBON is a weighting parameter and $S(a_i, b_j)$ reflects similarity of secondary structure marks. If the strict form of secondary structure description is used, then

$$Q(a_i, b_j) = 1 \text{ if } (T(a_i) = T(b_j) = \text{'H'}) \text{ or } (T(a_i) = T(b_j) = \text{'E'})$$
$$= 0 \text{ otherwise}$$

If propensity form is in use,

$$Q(a_i, b_j) = Hp1(a_i)*Hp2(b_j)+Ep1(a_i)*Ep2(b_j),$$

where Hp1, Hp2 are helix probabilities for 1st and 2nd sequences respectively; Ep1, Ep2 are strand probabilities for 1st and 2nd sequences. The "structure SW" score of alignment differs from the SW one only in one point: it uses the value $W(a_i, b_j)$ where SW score uses substitution score $M(a_i, b_j)$. Analogously, the STRUSWER algorithm, that finds an optimal alignment with respect to structure SW score can be obtained from the SW algorithm by calculating $W(a_i, b_j)$, where SW-algorithm calculates $M(a_i, b_j)$.

Alignment quality estimations. To compare algorithmic alignments with golden standard ones and to estimate quality of algorithmic alignment we use two characteristics, *accuracy* and *confidence*.

Alignment accuracy is the number I of positions *Identically* superimposed in algorithmic and golden standard alignment (GS) divided by the total number G of positions in the GS alignment.

$$Acc = I / G.$$

Alignment confidence is the number I of positions *Identically* superimposed in algorithmic and golden standard alignment divided by the total number A of aligned positions in the Algorithmic alignment

$$Conf = I/A.$$

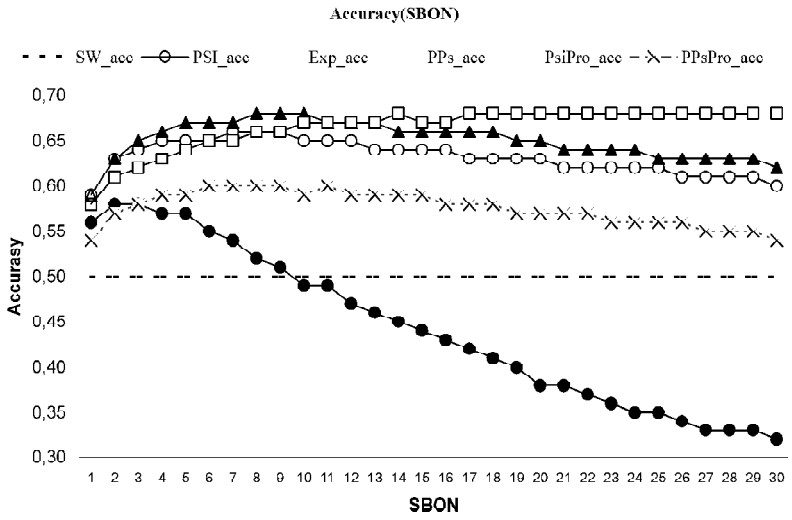


Fig. SBON parameter optimization. Methods abbreviations and description see in "Secondary structure" chapter. Each point represents average accuracy (given by Y-axes) of alignments obtained with a given SBON value (X-axes) and type of secondary structure assignment.

Results and Discussion

For every pair of proteins from the test set, we have constructed six alignments: the Smith-Waterman alignment (SW); and STRUSWER alignments with five types of secondary structure assignment (see "*Secondary structure*" above). In all cases we have used standard SW parameters, i.e. substitution matrix BLOSUM62, Gap Opening Penalty (GOP) 11, Gap Elongation Penalty (GEP) 1. To find optimal value of the parameter SBON we have tried all values from 1 to 30; and calculated alignment accuracy for all obtained alignments. The results are given in Fig. Table shows the average values of alignment accuracy and confidence, obtained for optimal values of SBON. The data are shown both for all data set, and for twilight zone only. One can see, that both experimental structures, and predicted by full version of PsiPred, allow significantly improve alignment accuracy without sacrificing with alignment confidence.

Table. Accuracy and confidence of various alignment methods (see notation in "*Secondary structure*"). The data are given for full test set (576 protein pairs) and for twilight zone only (protein pairs with sequence identity below 30 %, 368 pairs)

Method SBON	*SW* -	*Exp* 10	*PsiPro* 14	*Psi* 8	*PPsPro* 9	*PPs* 2
Full set (576 pairs) Acc	0,5	0,68	0,68	0,66	0,6	0,58
Full set (576 pairs) Conf	0,64	0,7	0,69	0,68	0,61	0,62
ID <30 (368 pairs) Acc	0,31	0,57	0,57	0,55	0,46	0,43
ID <30 (368 pairs) Conf	0,51	0,6	0,58	0,57	0,47	0,48

Aknowledgements

The work has been supported by the Russian Foundation for Basic Research (project Nos. 03-04-49469, 02-07-90412) and by grants from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003) and NWO.

References

1. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool // Mol. Biol. 1990. V. 215. P. 403–410.
2. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. 1997. V. 25. P. 3389–3402.
3. Jones D.T. Protein secondary structure prediction based on position-specific scoring matrices // J. Mol. Biol. 1999. V. 292. P. 195–202.
4. KabaschW., Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features // Biopolymers. 1983. V. 2(12). P. 2577–637.
5. Pearson W.R. Effective protein sequence comparison // Methods Enzymol. 1996. V. 266. P. 227–58.
6. Smith T.F., Waterman M.S. Identification of common molecular subsequences // J. Mol. Biol. 1981. V. 147. P. 195–197.
7. Sunyaev S.R., Bogopolsky G.A., Oleynikova N.V., Vlasov P.K., Finkelstein A.V., Roytberg M.A. From analysis of protein structural alignments toward a novel approach to align protein sequences // PROTEINS: Structure, Function, and Genetics. 2004. V. 54(3). P. 569–582.
8. Thompson J., Plewniak F., Poch O. BAliBASE: A benchmark alignments database for the evaluation of multiple sequence alignment programs // Bioinformatics. 1999. V. 15. P. 87–88.

AMINO ACID BIOSYNTHESIS ATTENUATION IN BACTERIA

***Lyubetsky V.A., Seliverstov A.V.* ***

Institute for Information Transmission Problems RAS, Moscow, Russia
* Corresponding author: e-mail: slvstv@iitp.ru

Keywords: *amino acid biosynthesis, attenuation regulation*

Summary

Motivation: Discovery and analysis of the "conventional" attenuation regulation in bacteria is currently an active field of research within the general framework of studying RNA-based regulation strategies, refer to (Vitreschak *et al*., in press) for details. However, finding alternative regulation systems is still a long way even in bacteria. Further, understanding the attenuation regulation machinery is crucial for developing algorithmic tools of mass attenuation detection and deriving a descriptive attenuation model. Identifying relevant attenuation characteristics to study is by itself a separate task.

Results: Putative attenuator structures in amino acid biosynthesis in Actinobacteria and *Staphylococcus aureus* are identified. Analysis of biosynthesis attenuation regulation in a wide range of proteobacteria and Gram-positive bacteria provided estimates of its characteristics.

Introduction

It is to be kept in mind that the "conventional" attenuation regulation of amino acid biosynthesis (this study operates with branched amino acids and leucyl-tRNA synthetase) implies presence of a leader peptide bearing regulatory codons (in fact, not only those encoding the amino acid), a terminator, antiterminator, a pause hairpin and a U-motif. Lyubetskaya *et al*. (2003) hypothesized that the hairpin formation requires a unique triplet word pattern, which was effectively implemented in the LLLM algorithm for mass detection of attenuation regulation (Gorbunov *et al*., 2001; for the search performance see Lyubetskaya *et al*., 2003; Vitreschak *et al*., in press).

Methods and Algorithms

All nucleotide sequences of leader regions as well as the gene annotations are obtained from NCBI. Conservative anchor motifs for use in multiple alignment were detected by our algorithm (Lyubetsky, Seliverstov, 2003). This algorithm involves the finding cliques in multipartite graph. It requires only polynomial time for computing a set of similar words in each nucleotide sequence.

Results

Actinobacteria. In many actinobacteria genes *ilvB*, *ilvN* (or *ilvH*) and *ilvC* comprise a single operon. The Table 1 shows the *ilvB*-containing operons' putative leader peptides, the operon type (second column) and the leader peptides' first nucleotide position as according to the NCBI nomenclature (third column).

Table 1. Leader peptides

Corynebacterium diphtheriae	*ilvBHC*	1081747
Met Asn **Ile Ile** Arg **Leu Val Val Ile** Thr Thr Arg Arg **Leu** Pro		
Corynebacterium efficiens YS-314	*ilvBHC*	1432212
Met Thr Ser **Ile** Arg Pro **Val Val Ile Val** Ala Ala Arg Arg **Leu** Pro		
Corynebacterium glutamicum ATCC 13032	*ilvBHC*	1337840
Met Thr **Ile Ile** Arg **Leu Val Val Val** Thr Ala Arg Arg **Leu** Pro		
Mycobacterium tuberculosis H37Rv	*ilvBNC*	3363125

Met **Leu Val Val Ile** Gly Arg Arg **Val** Gly Ala		
Mycobacterium bovis subsp. bovis AF2122/97	*ilvB-serA1*	3319743
Met **Leu Val Val Ile** Gly Arg Arg **Val** Gly Ala		
Mycobacterium leprae	*ilvBNC*	2046378
Met **Leu Val Val Ile** Cys Gln Arg **Val** Gly Gly		
Mycobacterium avium subsp. paratuberculosis str. k10	*ilvB1NC*	3381051
Met **Leu Val Val Ile** Arg Arg **Val** Gly Ala		
Mycobacterium marinum	*ilvB*	166742
Met Asp Thr Ala Gly Thr Pro Gly Lys **Leu Val Val Leu** Gly Arg Arg **Val Val** Ala		
Streptomyces avermitilis MA-4680	*ilvBNC*	3356481
Met Arg Thr Arg **Ile Leu Val Leu** Gly Lys Arg **Val** Gly		
Streptomyces coelicolor A3(2)	*ilvBNC*	6002909
Met Arg Thr Arg **Ile Leu Val Leu** Gly Lys Arg **Val** Gly		

Our alignment reveals that terminator hairpins and their preceding motifs are highly conservative in the organisms studied. Hereafter, the terminator half-stems are set in uppercase and the right-hand antiternimator parts are underlined:

C. diphtheriae cgaaaagcGCCCTCGaCAGCAccacacaTGCTGagCGGGGGCtttccttat

C. efficiens caagcGCCCTCGACAGTACccaccacaGTGCTGttTCGAGGGCtttgttgt

C. glutamicum caagcGCCCTCGaCAACACTcaccacAGTGTTGgaaCGAGGGCtttcttgtt

M. tuberculosis ccaacgcgACCCTCGtgCAGCagctgaGCTGgCGAGGGTtttttctt

M. bovis ccaacgcgACCCTCGtgCAGCagctgaGCTGgCGAGGGTtttttctt

M. leprae ccaacgcgcAACCCTCGtgCAGCTagtcAGCTGtCGAGGGTTttttgtt

M. avium ccaacgcgcAACCCTCGtgCAGCacaaGCTGtCGGGGGTTttttgtt

M. marinum ccaacgcgcAACCCTCGTgCAGCagctgaGCTGACGGGGGTTttttgtt

S. avermitilis cccggcgcgctCCCCTCGctTGCCtcacGGCACGAGGGGttttttgtt

S. coelicolor ccgacgcgctCCCCTCGctTGCCttacGGCACGAGGGGttttttgtt

In two actinobacteria, *Streptomyces avermitilis* and *Streptomyces coelicolor*, putative transcription attenuation regulation was found for a leucyl-tRNA synthetase gene *leuS* ortholog. The leader peptide: Met Arg Ala Val Arg **Leu Leu Leu** Ser Glu Pro Arg. Terminator hairpins are in uppercase, antiterminators underlined.

S. avermitilis (first nucleotide 6661741)

atgcgtgccgtacgc**cttctgctt**agcgagcc

gcgc**tga**tcagcccagaccactgacgattcgtggtcggaatcggcgcggcgtCCCCTCctgtgcGAGGGGttttttcatt

S. coelicolor (first nucleotide 2778624)

atgcgtgccgtacgc**cttctgctt**agcgagcc

gcgc**tga**tcagtcccgaccccggtcgtagtccggtggccggaatcggcgcggcgTCCCCTCctgtgcGAGGGGAtttttcatt

Staphylococcus aureus. The leader region of gene *ilvD,* which encodes dihydroxy-acid dehydratase in Gram-positive bacterium *S. aureus* contains a leader peptide preceded by a GA-rich SD region, a terminator hairpin with a U-rich motif and an antiterminator hairpin. The leader peptide possesses leucine and isoleucine codon strings:

Met **Leu** Asn Gln Tyr Thr Glu His Gln Pro Thr Thr Ser Asn **Ile Ile Ile Leu Leu** Tyr Ser **Leu** Gly **Leu** Glu Arg.

There are detected hairpins:

atg**ctt**aatcaatatactgaacatcaaccgacaacttcaaat**attattattttatta**tactct**tta**gga

ctcgaacgt**tag**taaatatttactaaacgctttaagtcctatttctgtttgaatgggacttgtAAACGTCCCAATAaTATTGGGACGTTTttttt

The first nucleotide position: N315 – 2097353; Mu50 – 2173855; MW2 – 2125745.

Table 2. Attenuation parameters

Bacteria	Operon	SU	L	G	C	100G/(G+C)
Actinobacteria						
Corynebacterium diphtheriae	*ilvBHC*	62	7	8	3	88
Corynebacterium efficiens	*ilvBHC*	69	8	7	3	70
Corynebacterium glutamicum	*ilvBHC*	67	6	8	2	80
Mycobacterium tuberculosis	*ilvBNC*	57	6	7	2	77
Mycobacterium bovis	*ilvB-serA1*	57	6	7	2	77
Mycobacterium leprae	*ilvBNC*	74	4	6	2	75
Mycobacterium avium	*ilvB*	72	4	7	2	77
Mycobacterium marinum	*ilvB*	59	6	7	2	77
Streptomyces avermitilis	*ilvBNC*	84	4	7	2	77
Streptomyces coelicolor	*ilvBNC*	84	4	7	2	77
Streptomyces avermitilis	*leuS*	66	6	5	0	100
Streptomyces coelicolor	*leuS*	70	6	5	0	100
Corynebacterium diphtheriae	*trpE1*	64	3	5	5	50
Streptomyces avermitilis	*trpE1*	47	3	5	3	62
Streptomyces avermitilis	*trpS2*	52	3	5	4	55
Staphylococcus						
Staphylococcus aureus	*ilvD*	78	1	4	1	80
Other						
Deinococcus radiodurans	*leuA2*	59	7	5	4	55
Deinococcus radiodurans	*ilvBN-x-C*	57	7	10	1	91
Thermus Thermophilus	*ilvBNC*	45	5	5	2	71
Bordetella	*thrS*	86	6	3	3	50
Ralstonia		78	2	4	3	57
Chromobacterium Vilaceum		51	2	3	4	43
Methylococcus capsulatus		101	1	4	2	66

Table 3. Earlier results for proteobacteria

	Operon	SU	L	G	C	100G/(G+C)	AS
alpha	*ilvIH*	51-55	4-7	2-5	1-3	40-66	
	trp(E/G)	52-72	3-10	4-6	2-5	44-66	
gamma	ilvBN	53-57	4-6	6-7	3	66-70	6 ..7
	ilvGMEDA	37-64	4-6	5-8	0-3	66-100	7 .. 31
	leuABCD	42-69	3-7	4-5	1-3	64-83	5 .. 33
	thrABC	46-62	3-8	3-7	1-3	50-88	-2 .. 22
	his	90-113	3-7	2-5	1-4	50-83	-6 .. 22
	trp	44-73	4-8	3-4	1-2	60-80	-8 .. -2
	pheA	61-72	3-5	4-6	1-2	71-86	-8 .. -6
	pheST	68-69	3-6	4-5	1	80-83	5 .. 33

The Tables 2 and 3 show some characteristics of transcription attenuation regulation. Data for proteobacteria are originally from (Vitreschak *et al*., in press). The third column contains distances SU between the initial position of the leader peptide stop codon and the beginning of the U-rich terminator hairpin region. The loop size of newly predicted terminators does not exceed 8, which well conforms to the known cases. The fifth and sixth columns contain the amount of G and C bases in the right half-stem of the terminator (preceding the poly-U). The distance AS between the antiterminator left half-stem and the stop codon varies between -8 (stop codon to the left of the antiterminator) and 33 (stop codon in the middle of the antiterminator loop).

Discussion

For *ilvB*-containing operons the distance SU is larger than in known proteobacteria. However, in operons *pheA*, *pheST* and *trp(E/G)* it is even larger and reaches 113 bases in *hisGDCBHAFI* operons. This parameter is a characteristic of the antiterminator structure properties. When the terminator hairpin is enough GC-rich, the proportion of Gs in its right half-stem is higher than that of Cs. The average ratio G/(G+C) = 2/3. The exception are some proteobacteria with a very short terminator containing nearly equal amounts of G and C (for instance, the right half-stem of the operon *ilvIH* terminator in *Rhodopseudomonas palustris* has a higher C content) and also low-GC bacteria. In the *ilvBNC* operon of actinobacteria predicted terminators possess a longer hairpin with the relative G content close to that in gamma-proteobacteria. Predictions in actinobacteria and other Gram-positive bacteria conform well to previous results. A stop codon can not be situated considerably far to the left from the antiterminator (rather, from the nucleotides complementary to a terminator hairpin region). The assumption AS > -9 seems to be strict. The number of regulatory codons in the leader peptide strongly correlates with the encoded amino acid. For tryptophan, a duplet or triplet of adjacent codons suffice. The *ilv* and *thr* operons involved in biosynthesis of several amino acids have leader peptides with numerous regulatory codons (14 codons preceding *ilvGMEDA* in *E. coli*). However, the *ilvIH* operon's leader peptide in alpha-proteobacteria contains between 3 (*Caulobacter crescentus*) and 6 regulatory codons. Predictions for *ilv* in actinobacteria and *S.aureus* contain at least 5 regulatory codons, which is congruent with evidence.

Acknowledgements

We wish to thank Dr. M.S. Gelfand for critical comments on this research.

References

Gorbunov K.Yu., Lyubetskaya E.V., Lyubetsky V.A. On two algorithms of detection of alternative elements of RNA secondary structure // Information processes. 2001. V. 1(2). P. 178–187.

Lyubetskaya E.V., Leontyev L.A., Lyubetsky V.A. Alternative secondary structure detection in a group of gamma-bacteria // Information processes. 2003. V. 3(1). P. 23–38.

Lyubetsky V.A., Seliverstov A.V. Selected algorithms related with finite groups // Information processes. 2003. V. 3(1). P. 39–46.

Vitreschak A.G., Lyubetskaya E.V., Shirsin M.A., Gelfand M.S., Lyubetsky V.A. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis // FEMS Microbiology Letters. 2004, (in press).

REPRESENTATION AND MODELLING OF PROTEIN SURFACE DETERMINANTS

Milanesi L.*[1], Merelli I.[1], Pattini L.[2], Cerutti S.[2]

[1] Institute of Advanced Biomedical Technologies, CNR, Segrate (Milan), Italy; [2] Department of Biomedical Engineering, Polytechnic University of Milan, Milan, Italy
***** Corresponding author: e-mail: luciano.milanesi@itb.cnr.it

Keywords: *protein surfaces, sequence comparison, site modelling, protein interface, homology*

Summary

Motivation: Surface characterization of peptides may provide useful information about functionality and potential interactions with other molecules. A description of a protein site through a surface that models the shape conferred by the exposed residues is an effective tool for the analysis and the modelling of proteins that may highlight similarities and relationships not detectable through comparisons at level of primary, secondary and tertiary structure.

Results: This study concerns the development of a tool that extracts the residues that concur to the shape modelling of the surface of a protein or a portion of it. This task is accomplished without taking into account the order of the amino acids in the primary structure, but only according to the selection of a portion of the protein indicated through geometric parameters or an explicit list of amino acids belonging to the site of interest.

Availability: luciano.milanesi@itb.cnr.it

Introduction

Much work has been carried out in the analysis of proteins at the level of primary, secondary and tertiary structure. Comparisons between different proteins performed on the basis of this kind of information have evidenced structural similarities that lead to important functional and evolutionary relationships. Moreover surface comparison makes negligible trivial matching of voluminous inner structures that are less informative in respect of interactions. A flexible surface description may allow the analysis of macromolecular complex interfaces to study protein-protein interactions. It is useful to segregate amino acids from a functional point of view: the identification of the *locus* of the residues that are exposed in a region of interest may constitute an additional information for the amino acid sequence. In this way protein comparisons can be performed on the basis of the information carried both in the primary structure and in the shape modelization leading to the definition of "superficial patterns".

Methods and Algorithms

The 3D atomic coordinates of a protein, as retrieved from the Protein Data Bank (PDB), are used to generate a so called *space-filling model* of it. This approach leads to an implicit modelization of the surface that is more suitable for this kind of analysis than the parametric one. Each atom is modelled as a volumetric item where is defined a function, which assumes negative values within van der Waals radius and positive values outwards, with a sign change just in correspondence of van der Waals surface. For example in the volume that contains an oxygen atom, sign inversion occurs at 0.73 Angstrom from the atom center. These items are positioned in a uniform space grid coherently with the PDB coordinates and are summed point to point. The sum allows to smooth the function values where sphere are contiguous or overlapped, avoiding the introduction of high

frequency noise. The resolution grid can be varied according the desired level of detail [1]. The tridimensional function defined on the grid is further low pass filtered with a filter box with 5 as size of the convolution kernel. This operation prevents the description of non representative details. The extraction of the surface is accomplished through the computation of an *isosurface* in the 3D grid function, which separates grid points with value below a defined threshold from those above. The linear interpolation is performed on the basis of the *marching cubes* principle which provides the triangulation of the surface. Note that to investigate the interaction of macromolecules it is important to examine their surfaces considering the solvent molecules which surround the Van der Waals surface in particular the protein modifying their accessibility. This is the Lee & Richards [2] model which is intuitively generated by rolling a probe sphere with a given radius (usually 1.4 Angstrom for water as the solvent since hydrogens are neglected) around the Van der Waals surface of the molecule: the trace of the center of the probe sphere produces the surface. A smoothed version of the Lee & Richards surface is calculated with the same algorithm used for the Van der Waals surface just modifying the sign inversion point in the volumetric item, through a specific threshold, that now occurs at the Van der Waals radius augmented by the solvent radius. The method is implemented in MATLAB which provides an ideal framework for both numeric analysis and graphics. The obtained mesh is straightforward visualizable, and many different options for the rendering are available, also the chance of coloring the surface according additional information, such as the electrostatic potential. As an example, see the rendered surfaces for FADH2 protein in Fig. 1.

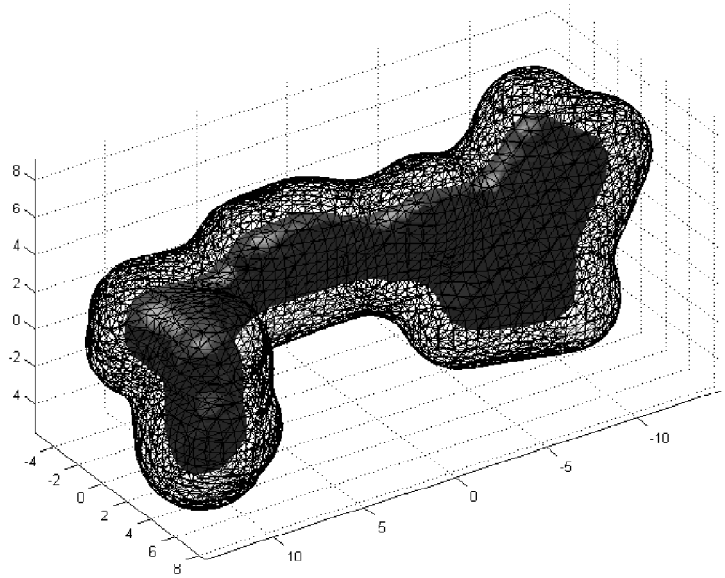


Fig. 1. Van der Waals (red) surface and solvent accessible surface (black) for $FADH_2$.

After surface extraction it is possible to isolate specific regions of it, that correspond to catalytical sites, characteristic domains or, in general, zones of interest. The selection can be realized by giving as input to the tool the list of the amino acids to be modelled or otherwise a portion of space described by geometric parameters. In Fig. 2 it is reported, as an example, the insulin protein on whose surface the active site that binds to receptors is evidenced and the detail of the mesh of the only site is depicted. Conversely, from the selected portion of a surface the subtended residues may be individuated. They are geometrically derived from the mesh that models the protein surfaces

at the interface. Then they are reported on the primary structure to enrich the information carried by the amino acid sequence. This allows comparisons between amino acid sequences on the basis of spatial relationships. It can be the case that two protein sequences are very different and the mere alignment does not show any significative similarity, but on the respective surfaces they reveal resembling patterns and functional key residues may be individuated. In this way internal structures, that are negligible if attention is on interaction phenomena, are not considered and only the amino acids that concur to the composition of the envelope are retained.

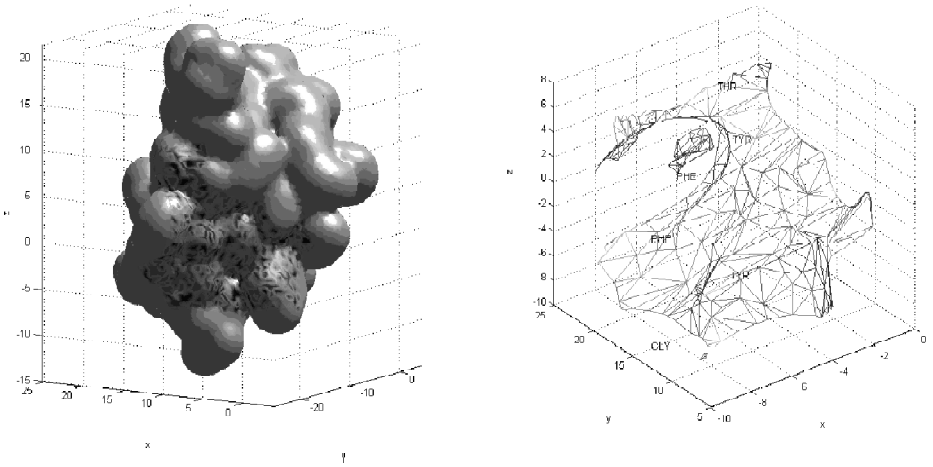


Fig. 2. Insulin protein: the active site that binds the receptors is evidenced (left) and the mesh of the only site is selected (right).

The portions of surfaces that constitute the interfaces of a few protein complexes have been considered. The attention is on the exposed residues that characterize the surface regions involved in the interactions. The first complex that has been considered is the Thrombin complex in *Homo sapiens* (PDB ID: 1a3e) and in *Bos taurus* (PDB ID: 1id5). This complex is constituted by an heavy chain (H), a light chain (L) and an inhibitor (I). Through the proposed method, on the heavy chain of both the organisms the trace of the interaction with the light chain can be evidenced, segregating only the residues exposed at this interface. There is a remarkable similarity of the selected patterns in the two species. The same procedure was followed to extract the amino acid of H involved in the interaction with I. Since I is different for the two complexes, its trace on H primary sequences does not show any common feature. Another example of interspecies comparison between surface determinants is shown for emoglobin, again a complex, which is constituted of two chains (A and B), is analyzed for human (PDB ID: 1bbb) and bovin (PDB ID: 1hda). The trace of the interaction of the chain B with chain A and the trace of chain A on chain B, respectively, are described: while a strong correspondence may be found between markers on A chains, it is clear that B chains are very dissimilar.

Discussion

The results presented are provided as an example of what kind of information can be retrieved through this approach. The developed procedure allows to describe portion of surfaces and identify putative functional residues that concur to the shape complementarity of peptides that interact. It is often the case that functionality is encoded in few residues and however it can be useful to visualize on the primary structure which amino acids are involved in the studied interaction. These

"surface patterns" so identified may lead to the identification of new proteins involved in a certain process and may help to better describe the mechanisms that govern interaction phenomena.

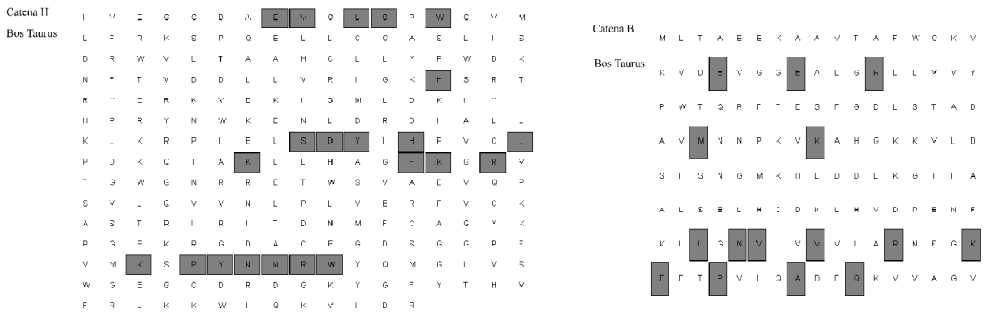


Fig. 3. Primary structure of thrombin (left) and emoglobin (right) where surfaces interacting amino acids are marked.

Acknowledgements

MIUR-CNR"Functional Genomics" 449/97. MIUR-FIRB "Bioinformatics" and "Enabling Platforms for high-performance computational grids orieinted to scalable virtual organizations" Project. MURST C.I.S.I. (Centre for Biomolecular Interdisciplinary Studies and Industrial Applications).

References

1. Vakser I.A., Matar O.G., Lam C.F. A systematic study of lowresolution recognition in protein-protein complexes // Proc. Natl Acd. Sci. 1999. V. 96. P. 8477–8482.
2. Lee B., Richards F.M. The Interpretation of Protein Structures: Estimation of Static Accessibility // J. Mol. Biol. 1971. V. 55. P. 379–400.

THE ALPHA-GALACTOSIDASE SUPERFAMILY: SEQUENCE BASED CLASSIFICATION OF ALPHA-GALACTOSIDASES AND RELATED GLYCOSIDASES

Naumoff D.G.

State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia,
e-mail: daniil_naumoff@yahoo.com

Keywords: *α-galactosidase, melibiase, glycoside hydrolase, GH-D clan, GH31 family, GHX family, COG1649, enzyme classification, protein family, protein phylogeny*

Summary

Motivation: About 1 % of genes in genomes code enzymes with glycosidase activities. On the basis of sequence similarity all known glycosidases have been classified into 90 families. In many cases proteins of different families have common evolution origin. It makes necessary to combine the corresponding families into a superfamily.

Results: Using of the PSI-BLAST program we found significant sequence similarity of several glycosidase families, two of which includes enzymes with the α galactosidase activity. Sequence homology, common catalytic mechanism, folding similarities, and composition of the active center allowed us to group three of these families – GH27, GH31, and GH36 – into the α-galactosidase superfamily. Phylogenetic analysis of this superfamily revealed polyphyletic origin of GH36 family, which could be divided into four families. Glycosidases of the α-galactosidase superfamily have a distant relationship with proteins belonging to families GH13, GH70, and GH77 of glycosidases, as well as with two families of predicted glycosidases.

Introduction

Glycoside hydrolases or glycosidases (EC 3.2.1.-) are a widespread group of enzymes, hydrolyzing the glycosidic bonds between two carbohydrates or between a carbohydrate and an aglycone moiety. A large multiplicity of these enzymes is a consequence of the extensive variety of their natural substrates: di-, oligo-, and polysaccharides.

Comparative analysis of 300 amino acid sequences of glycosidases known at the beginning of the 1990s showed that they could be classified into 36 families. Recent progress in genome sequencing resulted in collecting of a huge number of enzymatically-uncharacterized proteins: about 1 % of all genes encode enzymes with predicted glycosidase activities. Currently, more than ten thousand sequences of glycosidases and their homologues are known. They are grouped into 91 families: GH1 GH95 (except GH21, GH40, GH41, and GH60). Several glycosidases do not have any homologues. They are included into a group of non-classified glycoside hydrolases. Glycosidases catalyze hydrolysis of the glycosidic bond of their substrates via two general mechanisms, leading to either inversion or overall retention of the anomeric configuration at the cleavage point. Some related families of glycosidases, having the same molecular mechanism of hydrolyzing reaction, have been combined into clans. Currently, 14 clans (GH-A–GH-L) are described, and in total they contain 46 families (see Carbohydrate-Active Enzymes server, http://afmb.cnrs-mrs.fr/CAZY/).

Melibiases or α-galactosidases [E.C. 3.2.1.22] are glycosidases that cleave, with overall retention of the anomeric configuration, the terminal non-reducing α-D-galactose residues in α-D-galactosides, including galactose oligosaccharides, galactomannans, and galactolipids.

On the basis of sequence similarity, all α-galactosidases have been classified into four families of glycosidases: GH4, GH27, GH36, and GH57. Families GH4 and GH57 mostly include other glycosidases. The majority known б-galactosidases belong to GH27 and GH36 families which

form clan GH-D. Proteins of this clan have distant sequence similarity with representatives of several other families of glycosidases (Naumoff, 2001; Rigden, 2002). The recently established tertiary structure of several members of GH27 family is similar to the structure of retaining glycosidases from GH13 family (clan GH-H). Glycosidases of both families consist of the N-terminal catalytic $(\beta/\alpha)_8$-barrel domain and the C-terminal β-sandwich domain.

Results and Discussion

Sequences of the proteins belonging to the GH27 and GH36 families, according to the Carbohydrate-Active Enzymes classification, were used for BLAST screening of the GenPept database of amino acid sequences at NCBI server. The resulted database was enlarged by translation of nucleic acid sequences found by screening genomic sequences with the Genomic BLAST. In total we analyzed more than 300 proteins.

Family GH27 includes representatives from Eukaryota (Alveolata, Fungi, Metazoa, Mycetozoa, Viridiplantae) and Bacteria (Actinobacteria, Bacteroidetes, Fibrobacteres, Firmicutes, Proteobacteria). They possess the α-galactosidase, isomalto-dextranase [E.C. 3.2.1.94], α-N-acetylgalactosaminidase [E.C. 3.2.1.49], and galactosyltransferase [E.C. 2.4.1.] activities. Multiple sequence alignment of the full-length sequences of proteins from GH27 family shown that each protein has both domains characteristic of the family. Only three enzymatically-uncharacterized proteins contain solely the catalytic N-terminal domain. Some (mostly prokaryotic) proteins have additional domains, which we grouped into eight families by sequence homology.

Pairwise sequence comparisons showed that the majority of GH27 proteins have higher then 30 % identity, meeting the criterion of glycosidase subfamilies. All these proteins were grouped into 27a subfamily. Another subfamily, 27b, included five enzymatically-uncharacterized proteins from plants and bacteria. Two fungal proteins, including one α galactosidase, were considered to be the only representatives of subfamily 27c. A unique isomalto-dextranase from *Arthrobacter globiformis* and two other bacterial proteins do not belong to any of the subfamilies. The largest subfamily 27a included three subgroups, each containing sequences with no less then 50 % identity. The subgroups comprised proteins of yeasts, plants, and chordates, respectively.

Phylogenetic analysis of the GH27 family was used to study the evolutionary relationships of its members. Trees constructed by neighbor-joining and maximum parsimony methods (PHYLIP package) were topologically similar: all subfamilies (27a 27c) appear to form monophyletic groups with bootstrap value higher than 90 %. Eukaryotic proteins compose five distinct clusters of branches on the phylogenetic trees.

PSI-BLAST searches (*E*-value was 0.001 or 0.01) with a few randomly selected divergent representatives of the GH27 family used as a query sequence during the first or second iteration revealed some representatives of GH31 and GH36 families of glycosidases. The further iterations yielded members of GH-H clan (it includes families GH13, GH70, and GH77). Also we found a number of bacterial enzymatically-uncharacterized hypothetical proteins from several genome projects. Sequence analysis allowed to group them into two distinct families. One of them is known as COG1649. Another includes a unique α-glucosidase [E.C. 3.2.1.20] SusB from *Bacteroides thetaiotaomicron*, which belongs to the group of non-classified glycoside hydrolases. We have found the latter family for the first time and named it as the GHX family. Statistically significant similarity of GH27 glycosidases with members of the other protein families was only within the N-terminal catalytic $(\beta/\alpha)_8$-barrel type domain.

Families GH31 and GH36 includes representatives from Archaea, Bacteria, and Eukaryota. In addition to the α-galactosidase activity, α-N-acetylgalactosaminidase, stachyose synthase [E.C. 2.4.1.67], and raffinose synthase [E.C. 2.4.1.82] activities have been described for some members of GH36 family. Family GH31 includes retaining enzymes with α-glucosidase [E.C. 3.2.1.20], glucoamylase [E.C. 3.2.1.3], sucrase-isomaltase [E.C. 3.2.1.10 and E.C. 3.2.1.48], α-xylosidase

[E.C. 3.2.1.-], α-glucan lyase [E.C. 4.2.2.13], and isomaltosyltransferase [E.C. 2.4.1.] activities.

Multiple protein sequence alignment allowed us to find that two key Asp residues, playing the roles of nucleophile and proton donor in the enzyme active center, are located in the homologous sites of the catalytic domain in proteins of GH27, GH31, and GH36 families. Based on sequence homology, composition of the active center, common catalytic mechanism with overall retention of the α-D-glycopyranoside anomeric configuration of substrate during the reaction catalyzed, and predicted common $(\beta/\alpha)_8$ TIM barrel-type tertiary structure of the catalytic domain, we combined GH27, GH31, and GH36 families into the α-galactosidase superfamily (Fig.).

Phylogenetic analysis of proteins from the α-galactosidase superfamily showed that GH27 and GH31 appear to be monophyletic families and GH36 family is a polyphyletic one. Sequence analysis allowed us to distinguish in GH36 family four subgroups, which are monophyletic. We suggest considering these subgroups as four different families of glycosidases (GH36A-GH36D) belonging to the α-galactosidase superfamily (Fig.). Family GH36A includes proteins from Fungi and several phyla of Bacteria (Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria, Spirochaetes). Family GH36B contains only bacterial proteins (Actinobacteria, Proteobacteria, Spirochaetes, Thermotogales, Thermus). Among members of GH36A and GH36B families only the α-galactosidase activity has been shown. Family GH36C is composed by proteins from Archaea (Crenarchaeota), Bacteria (Actinobacteria, Bacteroidetes), and Eukaryota (Alveolata, Fungi, Viridiplantae). In addition to the α -galactosidase activity, stachyose and raffinose synthase activities have been described for this family. Family GH36D contains *Clostridium perfringens* α-N-acetylgalactosaminidase and a few enzymatically-uncharacterized proteins from Bacteria (Firmicutes, Proteobacteria).

On the basis of the available experimental data for SusB from *B. thetaiotaomicron*, protein sequence homology with glycosidases from the α-galactosidase superfamily, and the gene context we propose to consider the GHX family as a new family of glycosidases. However, taking into account a distant sequence similarity with α galactosidases and absence of experimental data about molecular mechanism and composition of the active center for the GHX family, at this point we have decided not to include the GHX family into the α-galactosidase superfamily. Statistically significant sequence similarity with glycosidases from the α-galactosidase superfamily and clan GH-H (which we propose to name the α-glucosidase superfamily) allows us to predict some glycosidase activities for proteins of COG1649 family.

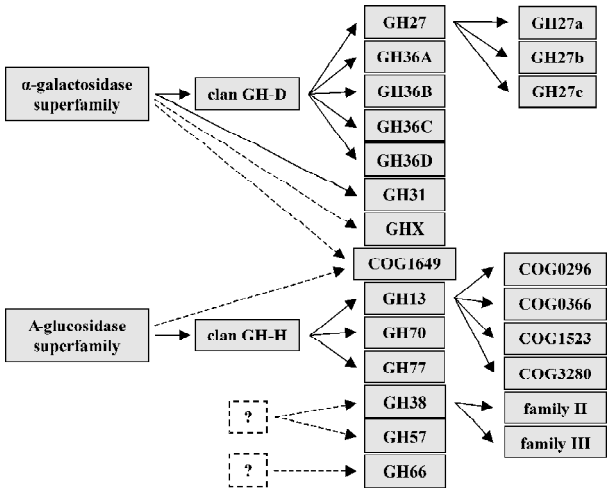


Fig. Proposed classification of $(\beta/\alpha)_8$-barrel-type retaining α-D-glycopyranosidases.

Our data strongly support a common evolution origin of proteins from the α galactosidase and α glucosidase superfamilies, as well as GHX and COG1649 families (Fig.). Homology of glycosidases from these two superfamilies have been proposed recently, as well as their distant relationship with some other retaining α-D-glycopyranosidase, representing families GH38, GH57, and GH66 (Henrissat, 1998; Imamura *et al*., 2001; Janeček, 1998; Rigden, 2002).
The results of this work including multiple sequence alignment and the updated classification of the α galactosidase superfamily may be obtained by e-mail.

Acknowledgements

This work was supported by a grant of the Russian President for young scientists (MK 118.2003.04).

References

Henrissat B. Glycosidase families // Biochem. Soc. Trans. 1998. V. 26. P. 153–156.

Imamura H., Fushinobu S., Jeon B.-S., Wakagi T., Matsuzawa H. Identification of the catalytic residue of *Thermococcus litoralis* 4-α-glucanotransferase through mechanism-based labeling // Biochem. 2001. V. 40. P. 12400–12406.

Janeček Š. Sequence of archaeal *Methanococcus jannaschii* α-amylase contains features of families 13 and 57 of glycosyl hydrolases: a trace of their common ancestor? // Folia Microbiol. 1998. V. 43. P. 123–128.

Naumoff D.G. Sequence analysis of glycosylhydrolases: β-fructosidase and a galactosidase superfamilies // Glycoconjugate J. 2001. V. 18. P. 109.

Rigden D.J. Iterative database searches demonstrate that glycoside hydrolase families 27, 31, 36 and 66 share a common evolutionary origin with family 13 // FEBS Lett. 2002. V. 523. P. 17–22.

INTER-SUBUNIT CONTACTS OF THE PROTEASOMAL ALPHA-SUBUNITS AS DETERMINANTS OF PARALOG GROUPS

Nikolaev S.V.*, Afonnikov D.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: nikolaev@bionet.nsc.ru

Keywords: *proteasome, proteasomal subunit, proteolytic complex, protein-protein interaction, mutual information, ortholog, paralog*

Summary

Motivation: Assignment of the proteasomal subunits to paralog groups has been initially based on phylogenetic analysis of amino acid sequences. The aim of the current work was to analyse the capability of the inter-subunit contact regions of the proteasomal subunits to determine the orderly arrangement of the alpha-subunits in the ring. This was because these regions are essential for the self-assembly of the proteasomes, and, hence, for the determination of the paralog groups.

Results. Based on the data on amino acid sequences and the known 3D structures of some proteasomes, we marked the positions of the alpha-subunits involved in inter-subunit contacts (the contact position, CP). There was a good match between the clustering based on the CPs and the one that was based on the complete sequences. Thus, we demonstrated that the CPs contain the information required for distinguishing the paralog groups.

Introduction

The proteasome is a multienzyme proteolytic machinery, which provides degradation of the bulk of cytoplasmic proteins up to oligopeptides. The proteasome is cylinder-shaped, composed of two identical halves, with each half arranged as two rings making up seven subunits each. The outer rings of the proteasome consist of the alpha-subunits, whereas the inner rings are composed of the beta-subunits. It is currently believed that the proteolytically active core of the proteasome (20S proteasome) results from the self-assembly of the subunits. The ring consisting of the alpha-subunits is self-assembled first, the ring composed of the beta-subunits is added after (Kopp *et al*., 1997). It is supposed that the orderly arrangement of the subunits in the proteasomal rings is stable. Therefore, it appeared of interest, what information on the alpha-subunit ordering in the ring is at the positions involved in the formation of the intersubunit contacts, because these positions indirectly participate in protein-protein recognition.

The accepted classification of the proteasomal alpha-subunits into paralog groups is based on phylogenetic analysis of amino acid sequences disregarding their structure and function (Bouzat *et al*., 2000).

The goals of this study were: 1) to clarify whether there is enough information to identify the alpha-subunit paralog groups at CPs; 2) to analyze the distribution of mutual information on the CPs.

Methods and Algorithms

Amino acid sequences of the proteasomal alpha-subunits given in Table 1 were extracted from the SWISS-Prot database, the data on the proteasomal 3D structures were extracted from the PDB.

The ClustalW program (Thompson *et al*., 1994) was used for multiple sequence alignment and sequence clustering. A clustering tree based on the results of the multiple alignment of the entire sample was used to set up groups of the alpha-subunit orthologs (Table).

To identify the contact positions, we regarded two amino acid residues as contacting if the distance between their C-alpha atoms was not greater than 6.5 Å. This threshold distance was used to

identify the sequence positions in the alpha-subunits in contact with any residue in any other alpha- or beta-subunit. Because each group of orthologs contained two subunits with known 3D structure, to identify the CPs, the union ("Union") of the CPs of an ortholog pair with known structure was first found, then, the intersection of the unions ("Intersection(Union)") was determined on all paralog groups. The thus chosen positions projected upon alignment sequences were accepted as the contact positions between paralogs.

Table. Analyzed sequences and clustering results obtained by the *ClustalW*

Paralog group	SWISS-Prot / PDB ID
1	1IRU:F, PSA1_HUMAN, PSA1_MOUSE, PSA1_RAT, PSA1_CHICK, PSA1_DICDI, 1FNT:F, PSA1_YEAST, PSA1_SCHPO, PSA1_CAEEL, PSA1_DROME, PS11_ARATH, PS12_ARATH, PSA1_ORYSA, PSA1_TRYBR, PSA1_TRYCR
2	1IRU:B, PSA2_HUMAN, PSA2_MOUSE, PSA2_RAT, PSA2_XENLA, PSA2_CARAU, PSA2_DROME, PSA2_ARATH, PSA2_ORYSA, 1FNT:B, PSA2_NEUCR, PSA2_SCHPO, PSA2_CAEEL, PSA2_TRYBB
3	1IRU:G, PSA3_HUMAN, PSA3_MOUSE, PSA3_RAT, PSA3_ARATH, PSA3_SPIOL, PSA3_ORYSA, PSA3_ACACA, PSA3_DICDI, PSA3_DROME, PSA3_CAEEL, 1FNT:G, PSA3_SCHPO
4	1IRU:C, PSA4_HUMAN, PSA4_MOUSE, PSA4_RAT, PSA4_DROME, PSA4_CAEEL, PS4L_DROME, 1FNT:C, PSA4_SCHPO, PSA4_ARATH, PSA4_SPIOL, PSA4_PETHY, PSA4_ORYSA, PSA4_DICDI
5	1IRU:E, PSA5_HUMAN, PSA5_MOUSE, PSA5_RAT, PSA5_DROME, PSA5_CAEEL, PS51_ARATH, PS52_ARATH, PSA5_SOYBN, PSA5_ORYSA, PSA5_SCHPO, 1FNT:E, PSA5_ENTHI, PSA5_TRYBB
6	1IRU:A, PSA6_HUMAN, PSA6_MOUSE, PSA6_DROME, PS61_ARATH, PS62_ARATH, PSA6_SOYBN, PSA6_TOBAC, PSA6_ORYSA, 1FNT:A, PSA6_YEAST, PSA6_SCHPO, PSA6_CAEEL
7	1IRU:D, PSA7_HUMAN, PSA7_MOUSE, PSA7_RAT, PSA7_CHICK, PS71_XENLA, PS72_XENLA, PSA7_CARAU, PS7L_HUMAN, PS7L_MOUSE, PS71_DROME, PS71_DROVI, PS72_DROME, PS71_ARATH, PS72_ARATH, PSA7_CICAR, PSA7_LYCES, PSA7_ORYSA, PSA7_DICDI, PSA7_CAEEL, PS73_DROME, PS73_DROVI, 1FNT:D, PSA7_YEAST, PSA7_SCHPO, PSA7_TRYBB

To evaluate the significance of a sequence position in the classification into paralog groups, we used mutual information at positions as a measure of the relation between an amino acid type at the position and a given paralog group: (Stormo *et al.*, 1986):

$$I_i = \sum_{x,y} f(x_i, y) \cdot \log \frac{f(x_i, y)}{f(x_i) \cdot f(y)},$$

where:

$f(x_i)$ – the frequency of amino acid type x at the i-th position in multiple alignment,

$f(y)$ – the proportion of proteins of the y-th paralog group,

$f(x_i, y)$ – the frequency of the amino acid type x at the i-th position in proteins of the y-th paralog group.

The RasMol (Sayle, Milner-White, 1995) package was used for visualiation, analysis and rendering of the protein structures.

Results and Discussion

Ordering of the alpha-subunits in the proteasomal ring. The proteasomal alpha-subunits extracted from the PDB (files PDB1FNT.ENT, PDB1IRU.ENT, Table) were used as markers of the alpha-subunits ordering in the ring. The clustering tree based on multiple alignment demonstrated a

subdivision of the chosen sequences with retention of non-intersecting classes of the marker subunits. A scheme for the ordering of the subunits in the alpha-ring of the proteasomes with known 3D structures and corresponding numbers of paralog groups, resulting from clustering based on multiple alignment, is given in Figure 1.

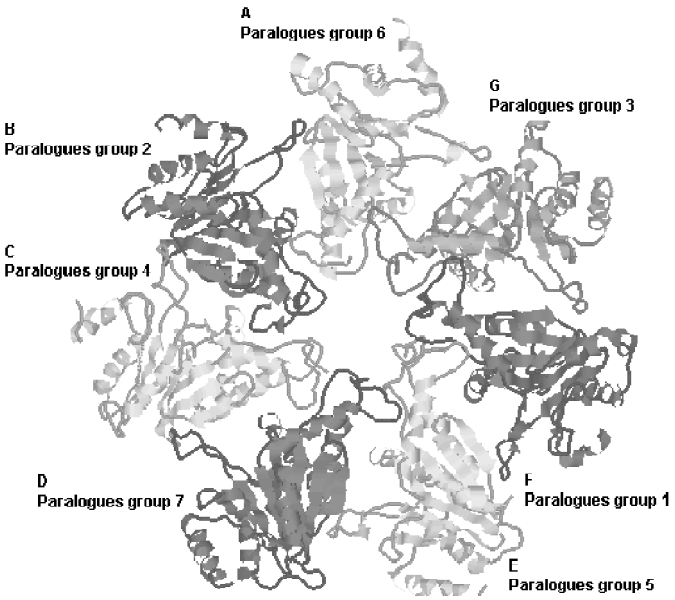


Fig. 1. The alpha-subunit ordering in the proteasomes with known structures (the one-letter code of proteasome subunit chains in files *pdb1fnt.ent*, *pdb1iru.ent* from the PDB) and paralog groups in which the subunits proved to be the clustering resulting from multiple alignment.

Clustering based on sequences composed of the amino acids involved in intersubunit contacts is similar to the one based on the entire sequences of the subunits. In contrast, when for the entire set of sequences, we took sequences consisting of contact positions chosen as Intersection (Union) and applied the ClustalW program, the clustering agreed well with the one obtained for the complete sequences. In fact, the paralog groups were the same as those listed in Table. This was evidence that the paralog group determinants were present at the contact positions chosen by the above method. Stating it in other words, the necessary conditions for distinguishing paralogs in protein-protein interactions at the alpha-subunit contact positions during the self-assembly fo the proteasomal alpha-ring are satisfied. However, the distribution of the mutual information at contact positions of a full-size sequence is identical. For this reason, it is an open issue whether these interactions suffice to provide a prepatterned orderly arrangement of the alpha-ring subunits.

Spatial structure of determinant distribution for the paralog groups at the contact positions. From the subdivision into paralog groups at multiple alignment of the alpha-subunits it already followed that the paralog groups differ by sequences. Consequently, it was of interest to localize the sites where the sequences were different. The question was: How significant were the contact positions in the distinguishment of the paralog groups?

Figure 2 shows the content of mutual information sufficient for differentiating the identity of paralogs at contact positions projected upon the proteasomal alpha-ring subunit. It is apparent that, among the positions involved in the inter-subunit contacts of the alpha-ring, certain are very strong determinants. Although the number of contact positions between an alpha-subunit and a beta-subunit was small compared with the one for the alpha-alpha subunits contacts, the former proved to be also strong determinants of paralog groups. This may be taken to mean that the alpha-beta contacts are important in ordering during self-assembly of the beta-ring in conjunction with the inter-beta subunit contacts.

The reference of an alpha-subunit to a paralog group depends on not only its position in the alpha-ring, but also on its position with respect to the beta-ring subunit.

Fig. 2. The content of mutual information for distinguishing paralogs at the contact positions projected upon the G-subunit of the proteasomal alpha-ring: a) a view along the proteasomal axis, b) a view of the inner surface of the alpha-ring. The residues with maximum content of mutual information are in black, those with minimal are in white.

Acknowledgements

This work was partly supported by grants from the Russian Foundation for Basic Research (03-07-96833-p2003, 03-07-96833); Siberian Branch of the Russian Academy of Sciences (projects No. 148,119); the Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501, subcontract 38/2004), RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4) and Origin and evolution of biosphere" (No. 10002-251/П-25/155-270/ 200404-082, the CRDF and the Ministry of Education of Russian Federation within the Basic Research and Higher Education Program (NO-008-X1, Y1-B-08-20).

References

Bouzat J.L., McNeil L.K., Robertson H.M., Solter L.F., Nixon J.E., Beever J.E., Gaskins H.R., Olsen G., Subramaniam S., Sogin M.L., Lewin H.S. Phylogenomic analysis of the á proteasome gene family from early-diverging eukaryots // J. Mol. Evol. 2000. V. 51. P. 532–543.

Kopp F., Hendil K.B., Dahlmann B., Kristensen P., Sobek A., Uerkvitz W. Subunit arrangement in the human 20S proteasome // Proc. Natl Acad. Sci. USA. 1997. V. 94. P. 2939–2944.

Sayle R., Milner-White E.J. RasMol: Biomolecular graphics for all // Trends in Biochemical Sciences (TIBS). 1995. V. 20. P.374.

Stormo G., Schneider T., Gold L. Quantitative analysis of the relationship between nucleotide sequence and functional activity // Nucl. Acids Res. 1986. V. 14. P. 6661–6679.

Thompson J.D., Higgins D.G., Gibson T.J. Clustal W. Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap-penalties and weight matrix choice // Nucl. Acids Res. 1994. V. 22. P. 4673–4680.

PROTEIN FAMILY PATTERNS BANK PROF_PAT. CURRENT STATUS

Nizolenko L.Ph.*[1], Bachinsky A.G.[1], Yarygin A.A.[1], Naumochkin A.N.[1], Grigorovich D.A.[2]

[1] SRC "Vector", Koltsovo, Novosibirsk, Russia; [2] Institute of Cytology and Genetics, Novosibirsk, Russia, e-mail: odip@bionet.nsc.ru

***** Corresponding author: e-mail: nizolenko@vector.nsc.ru

Keywords: *protein families, patterns, data banks, amino acid sequences, protein comparison*

Resume

Motivation: Protein family patterns bank Prof_Pat is one of the numerous "secondary" banks. Earlier we compared Prof_Pat to other known banks for an estimation of speed, sensitivity and specificity of the amino acid sequences analysis (Nizolenko *et al*., 2003). Now we compare it with Interpro (Mulder *et al*., 2003) to estimate completeness of the data, submitted in the Prof_Pat and recognising ability of the banks.

Results: From the 920.402 SWISS-PROT (rel. 42)\TrEMBL (rel. 25) sequences, that have the Interpro reference in their description, only 4 ones were no recognised by Prof_Pat patterns. At the same time, 14185 sequences, that have no this reference as well as any detailed description of a putative function for the protein, show very good similarity with well-described Prof_Pat patterns.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs//prof pat/ Prof_Pat local version is available via ftp: ftp://ftp.ebi.ac.uk/pub/databases/prof_pat/, ftp://ftp.bionet.nsc.ru/pub/biology/vector/prof_pat/.

Introduction

Now alongside with protein primary structure databases, the general recognition and a wide circulation have received so-called "secondary" databases in which the information on the whole groups (families) of the related proteins, most typical and frequently unique features of this group is concentrated. These bases are used for the analysis of amino acid sequences with the purpose of a prediction of functions and related communications of coded proteins. Bank Prof_Pat created and supported in the SRC VB "Vector" is "secondary" database too. We have compared it with Interpro, one of the largest and modern banks, which integrates databases UniProt, PROSITE, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIR Superfamily and SUPERFAMILY to estimate completeness of the data and recognising ability of the banks.

Methods and Algorithms

Protein family patterns, the bank of these patterns Prof_Pat and flexible fast search program were created using original technology (Bachinsky *et al*., 2000). The version of Prof_Pat 1.14 constructed on the basis of the 42th release of the SWISS-PROT bank and 25th release of TrEMBL contains patterns of 61973 groups of related proteins including more than 509500 amino acid sequences.

1.153.215 sequences from the SWISS-PROT (release 42) and TrEMBL (release 25) were used for the databases comparison.

Implementation

Prof_Pat is constantly updated database. The information on growth of this secondary base in parallel to growth of SWISS_PROT /TrEMBL volumes is submitted in Table 1.

To compare the recognising ability of Interpro and Prof_Pat banks, we have divided 1.153.215 sequences from the SWISS-PROT (release 42) and TrEMBL (release 25) into two groups: 1) entries, having the reference on Interpro in the description and 2) entries which have not such reference. Then we investigated these groups with Prof_Pat bank. The results of the analysis are presented in the Table 1.

Table 1. Growth of Prof_Pat in parallel to growth of SWISS_PROT /TrEMBL

Prof_Pat Release/data	SWISS-PROT /Trembl Release	Sequences in SWISS-PROT /Trembl	Sequences in Prof_Pat	Patterns in Prof_Pat	Protein Families in Prof_Pat
1.1 1998-99	29/1	~98000	52122	7083	
1.3 Feb 2000	38/11		~100000	~13000	
1.6 Oct 2000	39/15	295932	166667	24692	
1.7 Apr 2001	39/16	320511	181644	27187	
1.8 Nov 2001	40/17,18	385437	217360	31613	
1.9 Mar 2002	40/19	424908	248677	35917	
1.10 May 2002	40/20	475343	283765	41076	
1.11 Jan 2003	40/21,22	556538	344429	50149	40503
1.12 May 2003	41/23	634179	397627	57179	47533
1.13 Nov 2003	41/24	726011	462329	65331	55685
1.14 Jan 2004	42/25	784262	509506	71619	61973

Table 2. Comparison of the recognising ability of Interpro and Prof_Pat banks

SWISS-PROT(release 42) and TREMBL (release 25) sequences - 1.153.215		
Amino acid sequences, which **have** Interpro reference - 920.402		
Recognised by Prof_Pat families when release prepared	Recognised by Prof_Pat search program	No recognised by Prof_Pat patterns
572.149	348.249	4
Amino acid sequences, which **have not** Interpro reference - 232813		
Interpro reference as well as any detailed description is absent.	Recognised by well-described Prof_Pat families	Recognised by well-described Prof_Pat families with *Score*/m > 3
113.452	15.440	14.185

572.149 sequences from the first group have been found out in the files containing the description of Prof_Pat patterns. It means, that the bank during formation of new release has identified them.

The Prof_Pat search program studied the rest 348.249 sequences. At a level of similarity of 70 % only 4 from them appeared to be not identified. There were rather short (55–69 aa) fragments of amino acid sequences, 2 from which besides appeared to be hypothetical proteins.

The sequences which have no detailed description, but only "Hypothetical protein" or "ORF" in the DE field have been chosen from the second group. These sequences were studied by the Prof_Pat search program with the level of similarity of 90 %.

Families which description did not contain anything except for "Hypothetical protein" or "ORF" with identification number have been excluded from the protocol of comparison. Thus 15.440 sequences recognised by well-described Prof_Pat families were revealed. The founded similarity is certainly significant for 14.185 of them, because they have very high parameter *Score* used for an estimation of result quality. As it has been shown earlier (Nizolenko *et al.*, 2003), when Score/m (where m is the total number of motifs in the pattern) ≥ 3, more than 92 % false positive results are eliminated.

The majority of families identifying these sequences consist of some well-described proteins and 1-2 hypothetical proteins. But they have at least 30 % mutual similarity and very good alignment (Fig.). It allows asserting that hypothetical proteins included in family and their described relatives have the same function. Thus the bank appears to be the convenient tool in a prediction of possible function of the proteins not only at work with search system but also during formation of each new release.

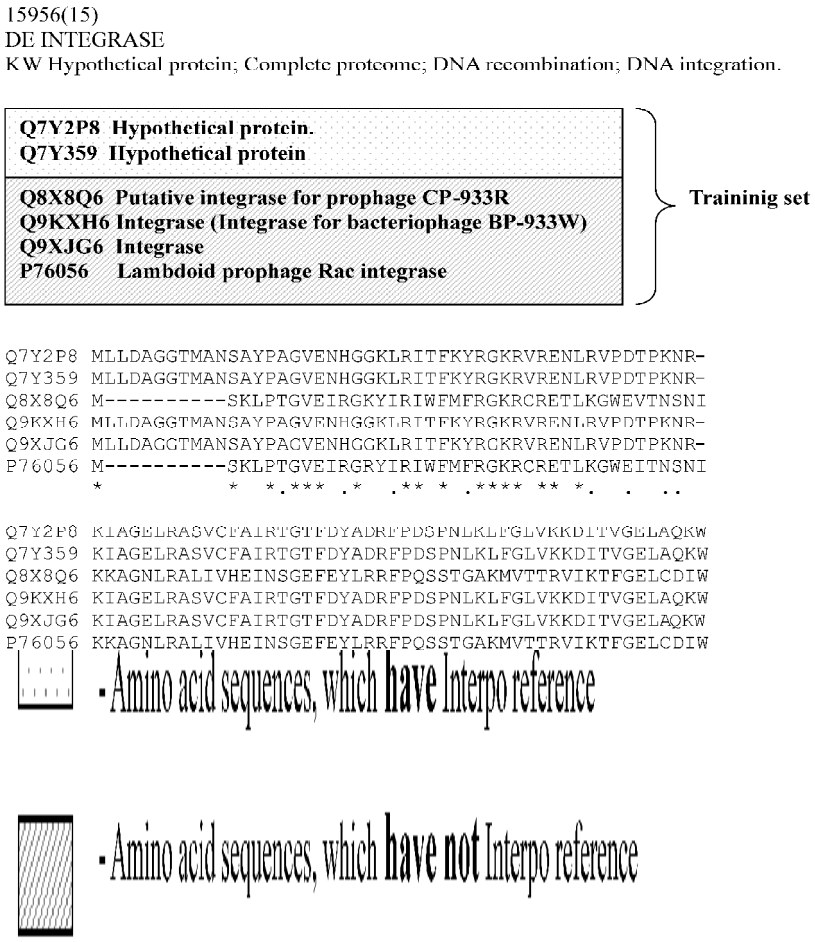


Fig. Fragment (the first 150 positions) of amino acid sequences alignment in PROF_PAT family 15956. This family includes well-described sequences as well as hypothetical ones.

Conclusion

The bank Prof_Pat identifies practically all sequences recognised by Interpro. At the same time with the help of Prof_Pat it is possible to describe rather significant number of the sequences which have no references on Interpro or other "secondary" databases.

References

Bachinsky A.G., Frolov A.N., Naumochkin A.N., Nizolenko L.Ph., Yarigin A.A. PROF_PAT 1.3: updated database of patterns used to detect local similarities // Bioinformatics. 2000. V. 16. P. 358–366.

Mulder N.J., Apweiler, Attwood T.K., Bairoch A., Barrell D., Bateman A., Binns D., Biswas M., Bradley P., Bork P., Bucher P., Copley R.R., Courcelle E., Das U., Durbin R., Falquet L., Fleischmann W., Griffiths-Jones S., Haft D., Harte N., Hulo N., Kahn D., Kanapin A., Krestyaninova M., Lopez R., Letunic I., Lonsdale D., Silventoinen V., Orchard S.E., Pagni M., Peyruc D., Ponting C.P., Selengut J.D., Servant F., Sigrist C.J.A., Vaughan R., Zdobnov E.M. The InterPro Database, 2003 brings increased coverage and new features // Nucl. Acids Res. 2003. V. 31. P. 315–318.

Nizolenko L.Ph., Bachinsky A.G., Naumochkin A.N., Yarigin A.A., Grigorovich D.A. Database of patterns PROF_PAT for detecting local similarities // In Silico Biology. 2003. V. 3. P. 205–213.

PROTEIN FOLDING AND MISFOLDING: A BIFURCATION STUDY OF A LATTICE MODEL

***Palyanov A.Yu.*[1], *Krivov S.V.*[2], *Titov I.I.*[1,3], *Karplus M.**[2,4], *Chekmarev S.F.**[1,5]**

[1] Novosibirsk State University, Novosibirsk, Russia; [2] Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 67000 Strasbourg, France; [3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia; [4] Department of Chemistry & Chemical Biology, Harvard University, Cambridge, MA 02138, USA; [5] Institute of Thermophysics, SB RAS, Novosibirsk, Russia
* Corresponding authors: e-mail: marci@tammy.harvard.edu, chekmare@itp.nsc.ru

Keywords: *folding and misfolding, kinetics, lattice heteropolymer, simulations, mutations*

Resume

Motivation: Protein misfolding is responsible for many human and animal diseases. An understanding of the mechanisms of folding and misfolding is essential to aid the search for inhibitors, which may prevent some of the misfolding diseases.

Results: A 27-residue heteropolymer on a cubic lattice, which has a compact stable ("latent") state competitive with the native state, is studied using Monte Carlo dynamics. The free energy surface of this system consists of three basins, which correspond to random semi-compact globule, native and latent states of the system. The region of bifurcation of folding pathways to the native and latent states has been determined, and the probability of folding into the native and latent states from various regions of the conformation space has been calculated. Structural memory effects are found to be present in the transitions between the native and latent states. The effect of the mutations on the folding time and probability of folding into the native state has been studied.

Introduction

The understanding of protein folding and misfolding mechanisms is one of the challenges of molecular biology. The availability of an ever-increasing number of genome sequences and the realization that misfolding is an important cause of disease, including Alzheimer's disease and spongiform encephalopathies, has greatly increased the effort to solve this problem. It is likely that the search for inhibitors of folding and aggregation, which may prevent some of the misfolding diseases, will be aided by an understanding of the mechanism of folding and misfolding. Considerable progress in resolving the folding problem has been made recently by a fruitful combination of theory and experiment (Dinner *et al.*, 2000), but there is still much to learn.

Methods

As the model system, we considered a 27-residue heteropolymer on a cubic lattice. Using a Go-like model, the monomer interaction energies were specified in such a way as to have two fully compact structures, which lowest (and equal) in energy but distinctively different in geometry. One of them, into which the system folds faster, is taken to be the native state of the system, and the other a competitive ("latent") state (Fig. 1). To simulate folding process, a Metropolis Monte Carlo (MC) algorithm was used. Three types of the moves were allowed; they are end flips, corner flips and two-bead crankshaft rotations.

Results

Starting folding trajectories with a randomly generated unfolded chain and terminating them upon reaching the native or latent state, we calculated the first-passage time of the system into these states. We found that in both cases, i.e. when the system folds into the native or latent states, the mean folding time exhibits a U-like variation with the temperature, with the minima achieved at T

= 0.6 and equal to approximately 3.9Ч10^5 MC steps. At the same time, the probabilities of attaining the native and latent states differ considerably; they are equal to 0.61 and 0.39, respectively.

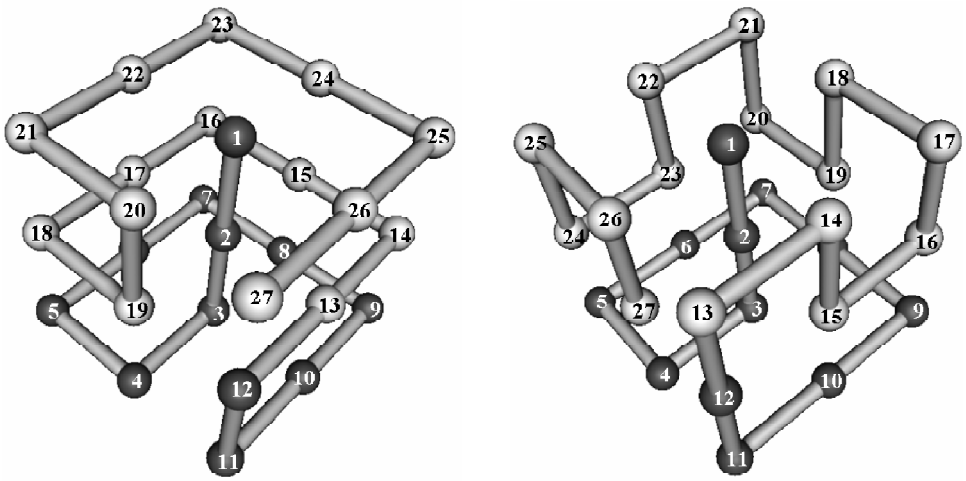


Fig. 1. Native (left panel) and latent (right panel) states of the heteropolymer. The part of the structures that is common for both states is shown in dark color.

Performing the equilibrium sampling of the conformation space, in which the trajectories were not terminated upon reaching the native or latent states but continued further exploration of the space, repeatedly visiting the native or latent states, we calculated the probabilities for the system to be in states with different numbers of total (N), native (N_{nat}) and latent (N_{lat}) contacts. These data were used to construct free energy surfaces as functions of N, N_{nat} and N_{lat}, with the free energy calculated as $F = -k_B T \ln P$, where T is the temperature, k_B the Boltzmann constant, and P the above mentioned probability. Figure 2a shows the free energy surface, depending on N_{nat} and N_{lat}, for the optimal folding temperature (T = 0.6), corresponding to the minimum value of the mean folding time. The surface consists of three broad basins; one of them corresponds to a semi-compact globule (small and moderate values of both N_{nat} and N_{lat}), and the other two to the native and latent structures (large values of N_{nat} and N_{lat}, respectively), with the energy biased towards these structures. Figure 2a shows that the native and latent structure basins are rather communicating through the semi-compact globule basin than directly connected. This inference is confirmed by calculated probabilities of transitions between the basins (Table). Structural memory effects are also found to be present in the transitions between the native and latent state basins (i.e., after leaving the native or latent state basin the system more frequently returns into the basin which it left).

Figure 2b presents the free energy surface for the same temperature as in Figure 2a, except that the sampling was not equilibrium; i.e., instead of a single long trajectory an ensemble of folding trajectories terminated upon reaching the native or latent state was considered. The surface differs from the equilibrium surface of Figure 2a mainly in that the native and latent states are no more lowest in free energy, because the system is not allowed to reside in the corresponding structures.

Figure 2 suggests that the folding paths undergo bifurcation at the boundary between the semi-compact globule basin and the native and latent state basins, with a higher probability to fold into the native state (0.61 against 0.39, as compared with the latent state). Figure 3 shows the probability of reaching the native state from various local regions of configuration space, characterized by different numbers of native and latent contacts. The dark color indicates the bifurcation points, from which the probability to fold into the native and latent states are approximately equal (0.45–0.55).

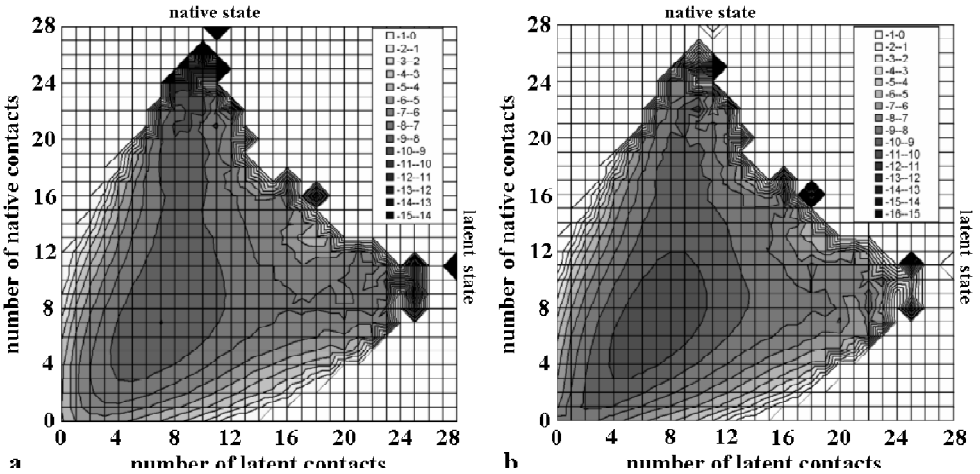


Fig. 2. Free energy surfaces depending on the numbers of native (N_{nat}) and latent (N_{lat}) contacts: (a) the equilibrium sampling, and (b) non-equilibrium (first-passage time) sampling. T = 0.6.

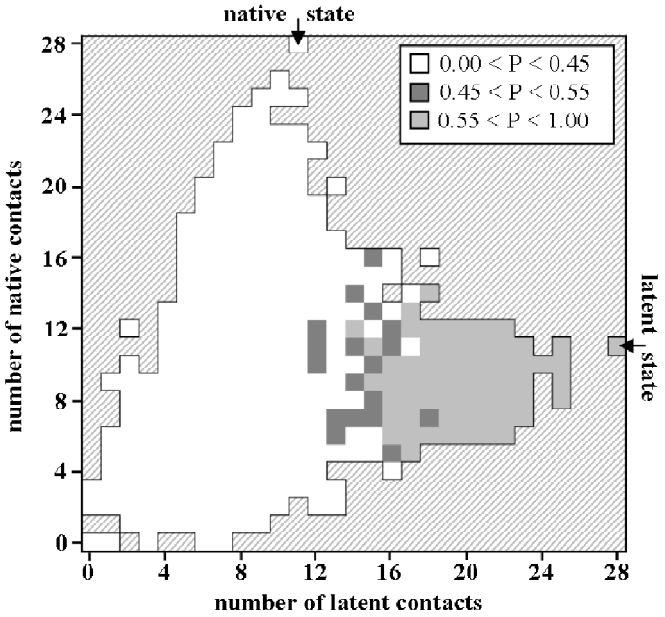


Fig. 3. The probability to reach the native state from different points of configuration space.

We also considered the effect of mutations on the mean folding time and probability to fold into the native state. The mutations were introduced by canceling the interaction in certain pairs of monomers. We found that mutations in the native contacts slightly increased the mean folding time and decreased the probability to fold into the native state, whereas the corresponding mutations in the latent contacts leaded to the inverse effect. Surprisingly, the mutations in the contacts common for the native and latent states, which practically did not change the probability to fold into the native state, leaded to a considerable increase of the mean folding time.

Table. Probabilities of transitions between the basins, including the intrabasin transitions (the diagonal elements of the Table), $T = 0.6$

From/To	Native	Semi-compact	Latent
Native	0.48	$0.95 \cdot 10^{-5}$	-
Semi-compact	$0.36 \cdot 10^{-5}$	0.18	$0.23 \cdot 10^{-5}$
Latent	-	$0.44 \cdot 10^{-5}$	0.34

Acknowledgements

This work was supported by the INTAS, grant #2001-2126. We also acknowledge a support from the RFBR, grant No. 2-03-32048 (A.P. and S. Ch.), SB RAS, grant Nos 119, 148 (A.P., I.T. and S. Ch.) RAS Presidium Program "Molecular and Cellular Biology" (project No. 10.4), Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501), and CRDF, grant #008-X1 (A.P. and I.T.). The work done at Harvard University was supported in part by the National Institutes of Health.

References

Dinner A.R., Šali A., Smith L.J., Dobson C.M., Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment // Trends Biochem. Sci. 2000. V. 25. P. 331–339.

STRUCTURAL MEMORY OF A LATTICE PROTEIN

***Palyanov A.Yu.*[1], *Titov I.I.**[1,2]**

[1] Novosibirsk State University, 630090 Novosibirsk, Russia; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: titov@bionet.nsc.ru

Keywords: *lattice heteropolymer, energy landscape, lifetime, folding, molten globule, generalized master equation*

Resume

Motivation: Brownian diffusion on energy landscape of few collective variables is a simple and conventional view of protein folding. As it is known from physics of disordered media, one can observe the non-Markovian kinetics when the landscape is corellated and/or additional thermodynamical variables are relaxing. Memory effects and its mechanisms for protein folding are still poorly explored.

Results: In this work we focused on kinetics of escapes from and returns to the native basin of a 27-mer lattice protein. We found that this basin contains 83664 compact and molten globule states with at least 15–16 out of 28 native contacts. The tail of basin lifetime distribution is Poisson with an exponent strongly differed from the inverse mean lifetime due to dominating of moves in the region of transition states. Once escaped from the basin the system returns back with overwhelming probability. In other words, the movement near the basin boundary is highly correlated what suggests a pronounced anisotropy of energy and\or connectivity of configuration space. Instead of differential equations of chemical kinetics, we describe the interbasin kinetics by the generalized master equation with memory.

Introduction

Protein folding is usually considered as stochastic movement on a landscape of collective variables, where native interactions control global downhill trend. Conflicts of large number of possible contacts between monomers create local energy traps on a landscape. When the populations of the minima are in inequilibrium the non-Markovian kinetics should be observed – well-known effect for a number of physical phenomena. A general physical reason of the anomaly is the relaxation of additional thermodynamical variable (Bouchaud, Georges, 1990); here even physical laws can change up to vanishing of transport coefficients (Yakobson, Titov, 1988). A trivial consequence of the effect is an initial stage of lattice protein folding. As the simulation starts, the probability of first passage to the native state grows, and only then kinetics turns to single-exponential (we don not consider low temperatures when the long-time kinetics may be stretched-exponential) (Lee *et al*., 2003). The growth stage is essentially determined by the initial condition and cannot be observed when equilibrium distribution was initially prepared.

Another general reason of non-Markovian relaxation of real proteins and their lattice models may be landscape correlations or irregularity of local connectivity. E.g., a single move of a lattice protein in physical space can form a number of contacts and will be observed as a ballistic trajectory in the reduced space of collective variables. Apparently, the anisotropy of energy landscape (local disconnectivity and energy valleys) define kinetic features of transition area beyond the basins, found in our work.

To examine non-Markovian dynamics of proteins below we replaced the master equation by the generalized master equation with memory after reduction to collective variables. Let us illustrate the mechanism of protein structural memory on a simplest example of the square lattice 6-mer. The chain has two compact states, corresponding to alpha- and beta-elements of secondary structure

(Fig. 1). Which contact forms first determines the folding outcome. Once the contact 2-5 is realized, beta-form will be more probable; the contact 1–4 (or 3–6) will preferably result in first passage of alpha-isomer.

Methods

We used the model protein from (Palyanov *et al*., 2004). Configuration space of this heteropolymer has two superbasins, native and latent; the dynamics of the protein is well-described by the following kinetic scheme (Palyanov *et al*., 2004).

Native state ↔ Semicompact globule and extended states ↔ Latent state.

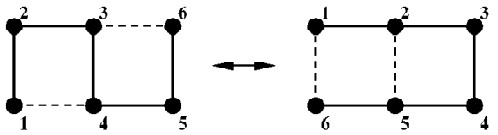


Fig. 1. Simple polymer with structural memory on a square lattice: 6-mer with two contacts in alpha (left) and beta (right) compact states.

Each return to the native basin was defined as an event, starting from the escape from the native basin and returning back without visiting the latent basin. To identify the basin states we modified the confinement technique (Chekmarev, Krivov, 1998) as follows.

The *confined enumeration* procedure searched for the states belonging to the basin. It has started from the native structure and has been consecutively run over nearest neighbors with the growth of free energy. When a boundary state was found, it became excluded for further visits. Totally we enumerated 83664 states in the native basin and 6943 states at its border. The latent states were enumerated in the same way.

The protein dynamics was simulated as usual: by Metropolis algorithm with three allowed types of monomer moves (crankshaft, end and corner flips) at optimal temperature of folding. The sampling trajectory spanned $3 \cdot 10^8$ Monte Carlo steps (mcs).

Results

First we will quantitatively characterize a molten globule of the native state. The rate of escape from the native basin exhibits Arrhenius dependence:

$$W \approx 300 \exp(-\frac{12.41}{kT}) \text{ (mcs)}^{-1}.$$

Here the activation energy is measured in units of monomer interaction energy. Hence the typical number of native contacts forming a molten globule can be estimated as at least 15–16 out of 28 total. And *vice versa*, it takes 12÷13 contacts to break for escaping from native basin. As we will see below, the basin escape is unexpectedly reversible.

Fast times dominated so strong in the basin lifetime distribution, that the mean lifetime ($3.9 \bullet 10^4$ mcs) was found about 10 times less than the characteristic time of Poisson tail ($4 \bullet 10^5$ mcs, Fig. 2a). Clearly the tail is responsible for the states close to the basin minimum. For further investigation of the movement in the basin neighborhood, we calculated the distribution of number of returns to the native basin (Fig. 2b). The distribution was found Poisson where the system typically exhibited a serial of 200 returns. The probability to change the basin of origin after escaping from the native basin was found very small, 0.015. A similar, though much weaker anisotropy of random walks was studied in details for persistent polymers and diffusion in solid states (best reviewed in Haus, Kehr, 1987). The correlation factor, accounting for the probability of an atom immediate return to the vacancy left behind after the jump, quantitatively characterizes this effect. The typical value of the correlation factor is about 0.6 depending on the lattice geometry. In our work we observed significantly stronger correlations with probability of returning back equal to 0.985. This result suggests a strong anisotropy of energy landscape beyond the basin. Both connectivity and energy valleys may be responsible for such an anisotropy (one can readily deduce this picture for 6-mer from Fig. 1).

Investigated dynamical features of the system are induced by correlated energy fluctuations and irregularity of the landscape connectivity, i.e. both by energetic and geometric disorder. Here the transition from microscopic to collective variables involves the transition to non-Markovian kinetics.

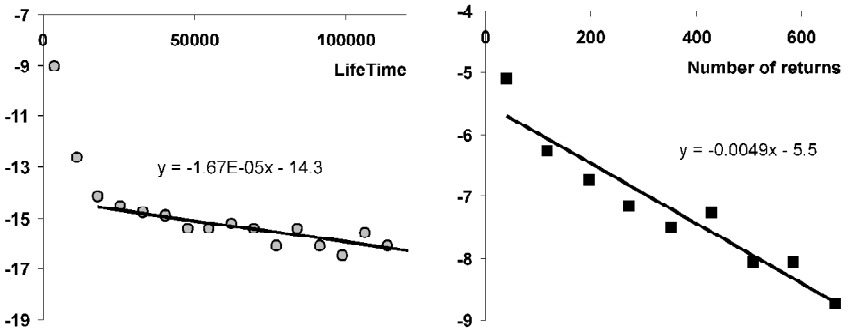


Fig. 2. The native basin escapes characterized by the basin lifetime distribution (left) and by the distribution of return number (right). The latter distribution was calculated as a number of visits to the native basin before a first passage of the latent basin happens. Both distributions are plotted in semi-logarithmic coordinates and shown are the linear regression equations. The qualitatively similar Poisson behavior was observed for the latent basin.

Let us follow, how it happens in terms of master equation describing the evolution of the probabilities of all particular states. If a coarse-grained variable enumerates a basin of location, summation over states in each basin leads to the generalized master equation for basin occupancies P_i:

$$\frac{\partial P_i}{\partial t} = \sum_{j \in \{i\}} \int_0^t W_m^{'}(t-\tau)[P_j(\tau) - P_i(\tau)]d\tau$$

with uniform memory kernels $W^{'}_{m}$.

Acknowledgements

The valuable criticism of S.F. Chekmarev on confinement enumeration is greatly acknowledged. This work was supported by a grant from the INTAS (#2001-2126), RFBR (# 02-03-32048), SB RAS (#119), Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”, Russian Ministry of Industry, Science, and Technologies (Nos. 43.073.1.1.1501). The research described here was made possible in part by Award No. NO-008-X1 of the U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF).

References

Bouchaud J.-P., Georges A. Anomalous diffusion in disordered media: statistical mechanisms, models and physical implications // Phys. Rep. 1990. V. 195(5/6). P. 127–293.

Chekmarev S.F., Krivov S.V. // Chem. Phys. Lett. 1998. V. 287. P. 719.

Haus J.W., Kehr K.W. Diffusion in regular and disordered media // Phys. Rep. 1987. V. 150(4/5). P. 263–406.

Lee C.-L., Stell G., Wang J. First-passage time distribution and non-markovian diffusion dynamics of protein folding // J. Chem. Phys. 2003. V. 118. P. 959.

Palyanov A.Yu., Krivov S.V., Titov I.I., Karplus M., Chekmarev S.F. Protein folding and misfolding: a bifurcation study of a lattice model (this issue). 2004.

Yakobson B.I., Titov I.I. Diffusion kinetics in the media with screw dislocations // Latv. J. Phys-Tech. Sci. 1988. V. 1. P. 88-90.

COMPUTER SIMULATIONS OF ANIONIC UNSATURATED LIPID BILAYER: A BASE SYSTEM TO STUDY PEPTIDE-MEMBRANE INTERACTIONS

***Polyansky A.A.*[1,2], *Volynsky P.E.*[2], *Efremov R.G.*[2]**

[1] Biological Department, M.V. Lomonosov Moscow State University, Moscow 119992, Russia;
[2] M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, 117997, Russia
***** Corresponding author: e-mail: newant@nmr.ru

Keywords: *molecular dynamics, DOPS, lipid-water interface, electrostatic interactions*

Summary

Motivation: Anionic bilayers represent a suitable model system to study peptide-membrane interactions. Structural properties of such systems have been extensively investigated in experiments. Nowadays, considerable new insight into the molecular mechanisms of proteins binding to membranes may be gained using computer simulations. This gives an opportunity to test theoretical predictions against direct experimental data. Molecular dynamics (MD) of peptides and proteins in full-atom hydrated lipid bilayers is one of the most powerful method to address these items *in silico*. Therefore, elaboration of physically reliable membrane-mimic systems and efficient computational protocols for their MD simulations is believed to be very important.

Results: A full-atom model of anionic unsaturated lipid bilayer surrounded with explicit water was created. The model contained 128 molecules of 1,2-dioleoyl-sn-glycero-3-phosphoserine (DOPS), the same number of positively charged Na^+ counterions, and about 6×10^3 waters. The system was subjected to 15-ns MD simulations with different algorithms of treatment of electrostatic interactions (cutoff function and particle-mesh Ewald summation (PME)). It was shown that under certain conditions, a good agreement between the macroscopic averages calculated over the equilibrium part of the MD trajectory and those available from experiments on DOPS bilayers, was achieved. On the other hand, in some simulations, the results demonstrated the artefacts, which dependent on the way of treatment the long-range electrostatics. Possible reasons for appearance of such artefacts are discussed.

Introduction

Protein-membrane and peptide-membrane interactions occur in many important processes in the living cell. That is why, such interactions attract growing interest in the field of structural biology. Although in recent years a large number of experimental approaches have been used to study the mechanisms of proteins' binding to membranes, it became evident that considerable insight into the problem may be provided by computer simulations. Particularly, MD represents a powerful tool to assess structural and dynamic parameters of model lipid bilayers.

Numerous studies of model zwitterionic bilayers containing the phosphatidylcholine group appeared during the last decade. At the same time, these are the anionic bilayers, which attract the major interest. They are similar to real biological membranes – almost all of them carry a negative charge on their extracellular surfaces. Phosphatidylserines are the most spread and biologically important lipids which can form such kind of bilayers. Unsaturated phosphatidylserine bilayers are commonly used in experimental studies of peptides and proteins in charged membranes. It is also important that such lipid bilayers present in the "native-like" liquid-crystalline state in extended range of temperatures. Unfortunately, wide application of computer simulation techniques to anionic unsaturated lipid bilayers is limited because of serious technical problems. They are mainly related to correct description of long-range electrostatic effects. Here we preset a new model system – the

hydrated DOPS bilayer with Na^+ counterions. Dynamical behavior of the system is explored *via* MD simulations with different computational protocols, and the one, which provides the best agreement of MD results with the experimental data [1], is selected for future applications.

Methods

MD simulations were carried out in the NPT ensemble – at T = 300 K and isotropic pressure of 1 bar. Electrostatic interactions were treated in two different ways: using 2-nm spherical cutoff, and the PME algorithm. The calculations were done with the help of the GROMACS v. 3.14 software [2]. MD-data processing was carried out using the GROMACS utilities and the set of auxiliary programs, which were specially written for this.

Results and Discussion

Analysis of the two accumulated MD-trajectories (15 ns each) permits the following conclusions. In both simulations the system retains well the bilayer conformation. Employment of the cutoff has larger effects on fluctuations of the bilayer thickness and geometry of the lipid-water interface, as compared to PME-calculations. Several important macroscopic averages estimated for the equilibrium part of both trajectories were compared with the experimental data (Table). It is seen that the order parameters for acyl chains, electron density profile, and bilayer thickness depend on the algorithm used to treat electrostatics. Moreover, analysis of radial distribution functions obtained for various sets of atoms and energy of electrostatic interactions permits delineation of differences in the structure of the water-lipid interface (data not shown). An important role of Na^+ interactions with polar headgroups of DOPS was proposed to explain such differences. The results obtained lead to a general conclusion that the PME scheme performs rather better (has higher correlation with experiments) than the cutoff-based protocol. In the last case, a number of artefacts in structural organization of the bilayer were observed. All of them were attributed to inadequate treatment of the membrane electrostatics. The model of DOPS bilayer that agrees well with the experimental data was selected for future work. It is currently being used to study interactions of peptides and proteins with anionic unsaturated lipid membranes.

Table. Equilibrium macroscopic averages of the DOPS bilayer: MD and experimental data

Parameter	Cutoff (2 nm)	PME	Experiment [1]
Order parameter for acyl chain, S_{cd}	0.10 ± 0.01	0.14 ± 0.01	0.15
Bilayer thickness, D_{pp}[1] (in nm)	3.8 ± 0.1	3.9 ± 0.2	3.9
Bilayer thickness, $L_{P\text{-}P}$[2] (in nm)	4.4	3.9	3.9
Area per lipid, Å^2	63.3 ± 0.1	63.3 ± 0.1	64.1

[1] D_{pp}, distance between P-atoms in different monolayers; [2] $L_{P\text{-}P}$ stance between peak of electron density function for P-atoms).

Acknowledgements

This work was supported in part by the Programme RAS MCB and the Ministry of Science and Technology of Russian Federation (the State contract № 43.073.1.1.1508), and by the Russian Foundation for Basic Research (grant 04-04-48875a).

References

1. Petrache H.I., Tristram-Nagle S., Gawrisch K., Harries D., Parsegian V.A., Nagle J.F. Structure and fluctuations of charged phosphatidylserine bilayers in the absence of salt // Biophys. J. 2004. V. 86. P. 1574–1586.
2. Berendsen H.J.C., van der Spoel D., van Drunen R. GROMACS // Comp. Phys. Comm. 1995. V. 91. P. 43–56.

ON AVERAGE ENERGY OF RANDOM WALKS WITH CONSTRAINTS AND GEOMETRICAL COMPLEXITY OF POLYMERS

Perevalov D.S.*[1], Davydov O.M.[2], Tatur S.V.[3], Lenskiy S.V.[3]

[1] Institute of Mathematics and Mechanics Urals Branch of RAS, Yekaterinburg, Russia; [2] Chelyabinsk's State University, Chelyabinsk, Russia; [3] Urals State University, Yekaterinburg, Russia
***** Corresponding author: e-mail: denis.perevalov@mail.ru

Keywords: *polymer, primary structure, random walk, energy, characteristic*

Summary

Motivation: The primary structure of the natural polymers (proteins, DNA, RNA) may be optimal in respect to some energetical characteristic of the set of the all possible polymer conformations.

Results: The characteristic of primary polymer structure which is based on Simon's definition of polygon's energy is proposed. We set as an example its usage analysing a class of sequences that contain equal-length blocks with two types of monomers. In this example the characteristic has one minimum, which is achieved on the sequence consisting of blocks with length 4.

Availability: http://www.biobase.ru

Introduction

When analyzing the functional properties of polymers, the problem of 3D structure analysis often arises. But in the majority of cases only the one-dimensional primary chemical structure is known. After this structure the geometrical 3D-structure is reconstructed not in a unique way. It may be useful to operate with the set of all the possible polymer conformations and to construct its energetic characteristics. One of such characteristics is proposed and the results of using it for analysis of the model class of sequences is given.

Model

We use the "random walks with constraints" approach for modeling a polymer molecule as a polygonal curve with the constraints depending on the sequence of monomers comprising the polymer.

Let us fix polymer sequence S and generate a sufficiently large number N of random polygonal curves c_i, $i = \overline{1,N}$ satisfying the constraints which depend on the sequence. Then for c_i each compute Simon's energy $E(c_i)$. It is defined as a sum of all $l(X)l(Y)/dist(X,Y)^2$, where X and Y are non-consecutive segments. Here $l(X)$, $l(Y)$ are the lengths of the segments and $dist(X,Y)$ is the distance between X and Y.

The energy is a scale-invariant and its value tends to infinity as the distance between two non-consecutive segments of the curve tends to zero. When the number of the curve segments is less or equal to 3, the energy is equal 0 because there are no non-consecutive segments. It can be proved that if the number of the segments is greater than 3 then the energetical minimum in the class of not-closed curves is achieved on the curves forming a straight line. Originally this energy was introduced as a characteristic of the polygonal curves entangling (Simon, 1994). In our case it is appropriate for analysis of the curve entanglement.

Define AE(S) as the average energy of the generated curves:

$$AE(S) = \frac{1}{N}\sum_{1}^{N} E(c_i).$$

This is the characteristic we propose. Actually $AE(S)$ characterizes the average entangling of $\{c_i\}_{i=1}^{N}$ or in other words "average geometrical complexity" of the sequence *S*. In our numerical experiments the distributions of the energy were unimodal. Thus we suppose, but can not prove, that $AE(S)$ is a good approximation of average energy of all random curves corresponding to *S*.

It should be noted that when the curve tends to have intersections, the energy of curve tends to infinity. But the huge value of some $E(c_i)$ may distort value of $AE(S)$ undesirable. To avoid the problem, we reject the curves having energy greater than some threshold. The threshold is obtained from experiments. Its value should be sufficiently large so that the main part of distribution lies below of the value.

Having computed the average energy for several sequences, one can compare their geometrical complexity and find energetically the optimal sequence.

Results and Discussion

We have applied this definition of average energy to analyse a simple class of polymers. We used Freely Rotating Chain model, also known as Worm Like Chain model of polymers (Kratky, Porod, 1949). This model assumes that there are constraints on the valency angles of adjacent segments, but allows the free rotation about bond axes (i.e. torsional angles are random).

There are two types of the monomers denoted by symbols L and H in our model. Symbol L means that the valency angle should lie in $[120^o,135^o]$. Symbol *H* means that the valency angle should lie in $[60^o,90^o]$. The distribution of a valency angle in an appropriate angle segment is random. All lengths of bonds are equal. To construct a random polygonal curve corresponding to the sequence S we create the first node at arbitrary position and direction, then generate the other nodes using the constraints arising from *S*.

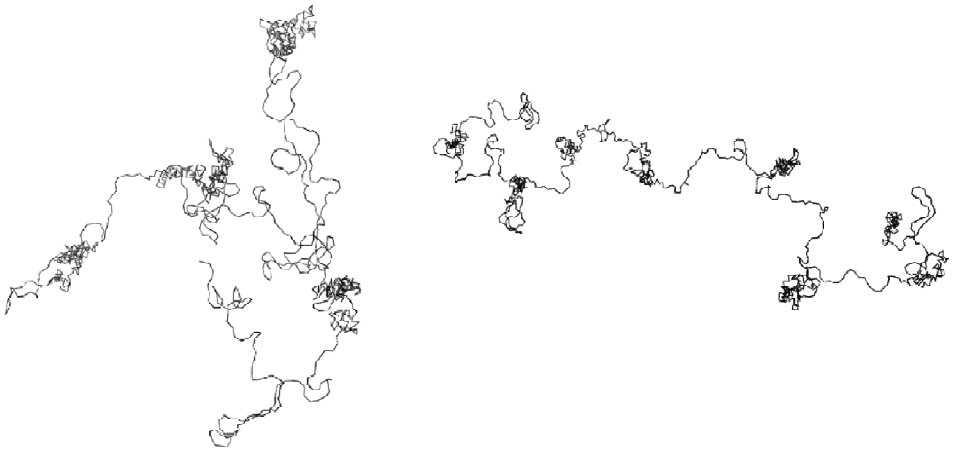


Fig. 1. Examples of random curves generated by sequences having 8 and 16 blocks.

We consider the class of sequences of size 1024 and consisting of successive equal-length blocks of *L*'s and *H*'s. For instance, two random curves corresponding to the sequences with 8 and 16 block are presented on Fig. 1. One can see how blocks of *H*'s in the sequences provide dense regions in the curves.

We have computed the average energy for a number of sequences having 8, 16, 32, 64, 128, 256, 512 and 1024 blocks (the number of used curves was 20000, energy threshold value was 100000). The Fig. 2 depicts the graph of dependence between the number of blocks and the average energy.

It unexpectedly appeared that the graph is non-monotonous. It gives the minimum value when the number of blocks is equal to 256. It means that in the considered class of sequences there is an energetical optimum provided by the sequence with the size of block equal to 1024/256 = 4. The significance of the average energy minimum presence is not clear now and is the theme of a further research.

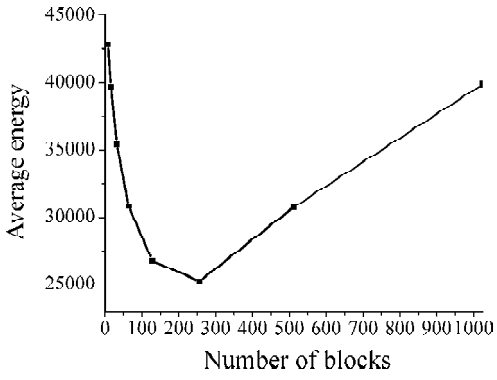


Fig. 2. Dependence between the number of blocks in the sequence and its average energy.

The following question arises from the obtained results. If we considered the sequence of real polymer monomers, would it be optimal in respect of an energetic characteristic like $AE(S)$?

Acknowledgements

We are very grateful to K.W. Gilbert and E.A. Biryukova for useful discussions and A.M. Davydova for invaluable help in translation.

References

Garcia de la Torre J. *et al*. Calculation of the solution properties of flexible macromolecules: methods and applications // Eur. Biophys. J. 2003. V. 32. P. 477–486.

Kratky O., Porod G. Rontgenuntersuchung Geloster Fadenmolekule // Recueil Trav. Chim. 1949. V. 68. P. 1106–1122.

Simon J.K. Energy functions for polygonal knots // J. Knot Theory and its Ramifications. 1994. V. 3. P. 299–320.

A MOLECULAR MECHANISM FOR THE STRUCTURE-FUNCTIONAL ALTERATIONS IN MUTANT FORMS OF HUMAN P53 PROTEIN

***Pintus S.S.*[1], *Ivanisenko V.A.*[2]**

[1] Novosibirsk State University, Novosibirsk, Russia; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

***** Corresponding author: e-mail: pintus@bionet.nsc.ru

Keywords: *gain-of-function mutations, p53, structure alignment, molecular dynamics*

Summary

Motivation: The exact molecular mechanisms of the effects of the gain-of-function mutations are not known as yet (Gualberto *et al.*, 1998). Here, we performed a computer analysis of the possible molecular mechanisms of the gain-of-function effect of cancer-prone mutations in human *p53*.

Results: In this study we identified new functional sites in mutant human p53 protein, using the structure alignment technique. Molecular dynamics simulations were used to demonstrate the efficiency of the found sites. We also suggested a molecular mechanism of the gain-of-function effect in the mutant p53 protein.

Introduction

The p53 protein is a key factor for cell cycle regulation and apoptosis in tumor cells. If DNA of a cell is damaged, the protein can stop the cell cycle in the G1 phase (Kastan *et al.*, 1991); this allows the cell to undergo reparation before replication starts or apoptosis that would result in cell death (Lowe *et al.*, 1993). The *p53* gene is mutated in 45–50 % of human tumor cells (Greenblatt *et al.*, 1994). The prevailing reason for cancer diseases is point substitutions in the *p53* gene. The substitutions are of particular interest because they often do not inactivate the *p53* product (Soussi *et al.*, 2000). The substitutions include the so-called gain-of-function mutations that result in new oncogenic properties of a mutant protein.

We proposed that some mutations in the *p53* gene may result in appearance of new functional sites. These sites determine the gain-of-function effect. To test this assumption, we searched for new functional sites in the mutant forms of the p53 protein.

Methods and Algorithms

The previously developed program, PDBSiteScan, was used to search for new functional sites in the p53 protein. Its algorithm is based on step-by-step extension of a set of juxtaposed residues of a predictor site and a protein using restricted search (Ivanisenko *et al.*, 2002). PDBSiteScan searches for active, binding and posttranslational modification protein sites; the program structurally aligns the 3d structure of a protein with every exemplary site stored in the PDBSite database to identify the site searched in the protein (Ivanisenko *et al.*, 2002).

To search for new sites in the mutant forms of the p53 protein, the data were processed as follows. The 3d structure of a normal protein was taken from the PDB database (Berman *et al.*, 2000); then, a normal residue was assigned a mutant name in the resName field for every ATOM record of a mutant residue. Thus, a new entry in the PDB format was obtained. This entry contains the atom coordinates for the normal protein and a primary sequence for the mutant. PDBSiteScan aligns only atoms of the main protein chain known to be not highly altered due to missense mutations (Soussi *et al.*, 2000). On the other hand, PDBSite demands identity of two primary sequences of the exemplary site and the one to be predicted. In this way, search for the functional sites in the tertiary structure of

a mutant protein using the PDBSiteScan program is implemented. The data on the *p53* gene mutations was obtained from the P04637 entry of the Swiss-Prot database (Boeckmann *et al.*, 2003).

The tertiary structures of the p53 mutants were calculated as follows. PDBSiteScan allows to obtain a PDB format file of the alignment of a potential site in a normal protein structure with the corresponding exemplary site from the PDBSite database. The obtained file was edited so that the Cartesian coordinates of the predicted site were replaced by those of the exemplary site. As a result, the new Cartesian coordinates of the N, C, and C_α atoms of the predicted functional site were obtained. The 1GZH, 1C26 and 1YCR structures were used for new site search. The Cartesian coordinates of the zinc ion were also extracted from the PDBSite database. They were subjected to the affine transformation derived from the site alignment. In this way, the zinc ion coordination in the mutant p53 structure was estimated.

Energy minimization of the obtained structures was used to evaluate the stability of the identified sites by the Swiss-Model server. Swiss-Model minimizes the energy of protein structures in the GROMOS96 force fields (Schwede *et al.*, 2003). The residues of the predicted site were supposed to acquire the conformation similar to that of the exemplary functional site taken from the PDBSite database.

Molecular dynamics simulations of the mutant protein structures were performed to test our assumtion on the functionality of the sites predicted using the GROMACS 3.1 (Lindahl *et al.*, 2001) program package.

Implementation and Results

New functional sites were searched in the DNA binding, tetramerization and transactivation domain structures of the human p53 protein using the PDBSiteScan program. The domains are represented in the PDB databank by the 1GZH, 1C26 and 1YCR crystal structures, respectively. The structures were modified as described above. The substitutions 175 R->H, 245 G->C and 245 G->D were analyzed. These substitutions have been shown to result in gain-of-function activities of the mutant p53 protein (Sigal *et al.*, 2000).

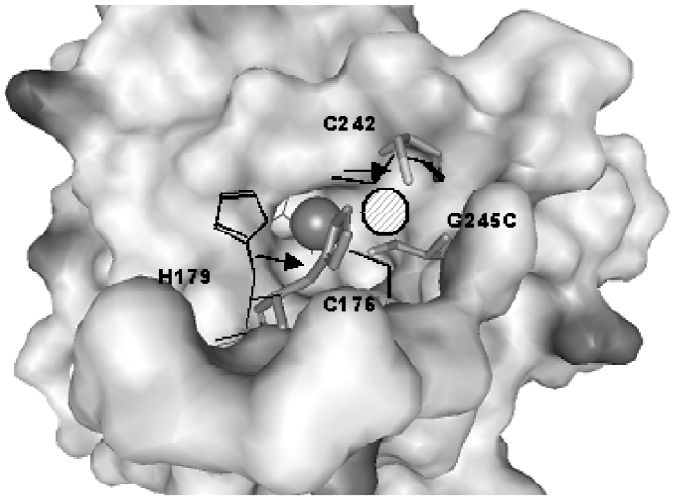


Fig. 1. The Zn^{2+} binding site of the DNA-binding domain of the human p53 protein (PDB Id 1gzh). Thin lines indicate the residues of the normal Zn^{2+} binding site of the p53 protein. The residues resulting from 245G->C mutation in additional Zn^{2+} binding site are shown in the sticks model. The structure of the new site was determined by structural alignment of this p53 protein fragment with the Zn^{2+} binding site of cytidine deaminase (PDB Id 1AF2). Arrows show the relocations of the residues to occupy a position at which the new site uptakes the Zn^{2+} ion. The Zn^{2+} ion the normal site uptakes is depicted by a ball, the one the potential site uptakes by a hatched circle. The image is obtained using the ViewerLite package.

The 245 G->C substitution was found to result in the formation of new zinc binding sites in the mutant p53 protein (Fig. 1).

The conformational stability of the potential sites identified in the 245G->C mutant form was demonstrated using the Swiss-Model server.

The normal and mutant proteins in complex with the zinc ion were solvated after preliminary energy minimization was modeled. The obtained structure of the p53 mutant protein, with the predicted Zn^{2+} ion position, and the structure of the normal p53 protein (PDBID 1TSR) was subjected to molecular dynamics using simulations for 10 ps in water solvent at pH 7.0 after preliminary energy minimization by the steepest descent using GROMACS 3.1.

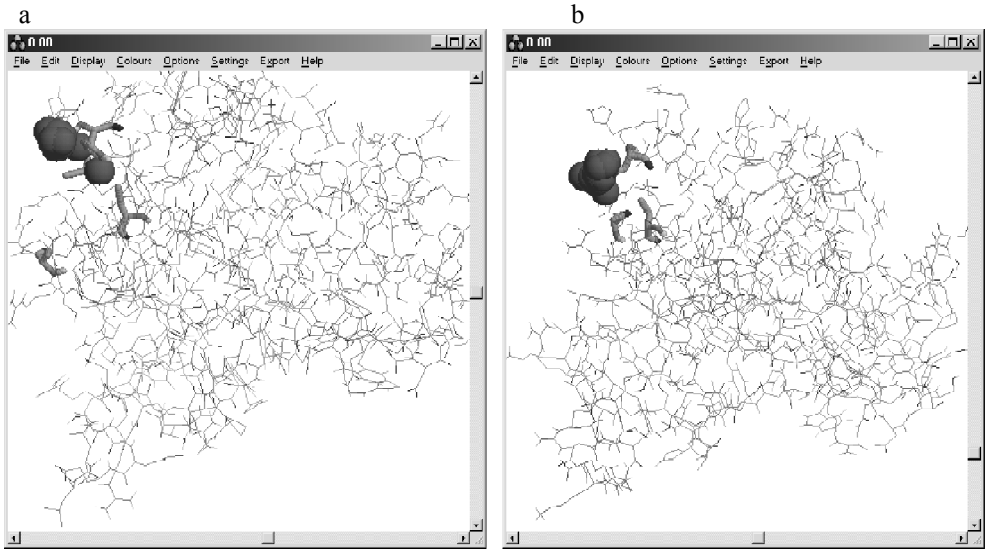


Fig. 2. The Zn^{2+} ion trajectory in the normal p53 protein (a) and the 245 G->C mutant (b). Balls indicate the position of the Zn^{2+} ion at diferent time points. The amino acids of the normal and new binding sites of the Zn^{2+} ion are depicted as sticks. The view was obtained using rasmol.

The energy of Coulomb and Van der Waals interactions of the Zn^{2+} ion with the normal and mutant forms of the p53 protein in solvent at different points of the trajectory with 0.2 ps per step was calculated. Modeling demonstrated that in the mutant p53 protein the Zn^{2+} ion displaces deep into the molecule to approximate the new potential functional site (Fig. 2).

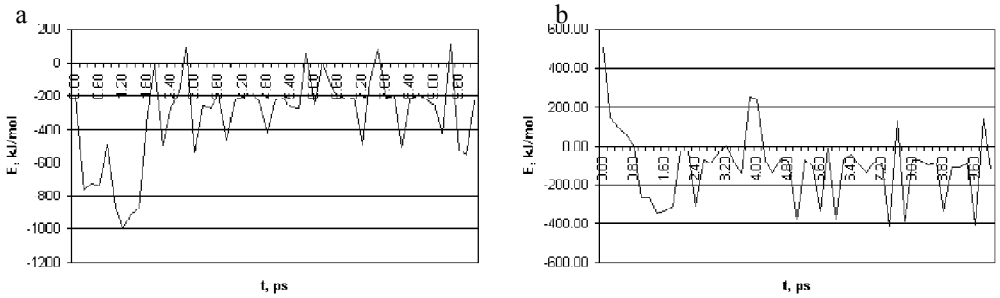


Fig. 3. The energy of Coulomb and Van der Waals interactions of the Zn2+ ion with the normal (a) and mutant (b) proteins.

Modeling allowed us to compare the dynamics of the interaction energy of the Zn^{2+} ion with the normal and mutant p53 protein (Fig. 3). Clearly, the binding patterns of the zinc ion in the normal functional site and the one we identified in the mutant form are different (Fig. 3).

In the case on the normal protein, total energy of Coulomb and Van der Waals interactions is maintained near its average value, -325.153 kJ/mol, whereas in the case of the mutant protein the energy decreases from 503.142 kJ/mol, approaching the average value –100.932 kJ/mol. This means that the zinc ion interacts with the functional site we predicted.

Discussion

The obtained results suggest the following molecular mechanism for the effect of the 245G->C substitution. The substitution results in a potential Zn^{2+} binding site. The site is positioned not far from the real and can compete for the zinc ion with it. This may be associated with change in the DNA-binding domain, thereby making the p53 protein express new genes. Experimental data have indicated that 245G->C substitution disrupted the spindle checkpoint and resulted in polyploidy in Colcemid-treated Li-Fraumeni fibroblasts. Subsequently it has been shown that the disruption may be due to the overexpression of the hsMAD1 gene when its promotor binds to the mutant p53 protein (Iwanaga, 2002).

It should be noted that all the amino acids referred to this sites reside in the same protein pocket as those of the real site. This makes ion transfer from one site to site to another quite plausible. The new site is positioned in the close vicinity to the existing Zn^{2+} binding site. It is known that the Zn^{2+} ion modulates binding the p53 protein to DNA, for this reason disruption of the function of the site binding to the zinc ion may affect the DNA binding function in the mutant p53 protein.

It is important to remember that all mutations we analyzed here occur both in families with Li-Fraumeni syndrome and in patients with colon cancer (Hollstein *et al.*, 1991). The mutant p53 protein can acquire new functions that the normal protein does not possess (Geutskens *et al.*, 2000). It is of interest that the colon cancer is not a feature of Li-Fraumeni syndrome (Levine *et al.*, 1993). Competition of the mutant site with the normal for the Zn^{2+} ion may be common to these two different cancers.

There appear to be good reasons for suggesting that the normal site and the one resulting from mutation compete for binding the Zn^{2+} ion.

Acknowledgements

The research described in this publication was made possible in part by Award No. REC-008 of the U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF). The work was supported by the Russian Foundation for Basic Research (02-07-90335, 03-07-96883-p2003 03-07-96833, 03-07-90181-b, 02-04-48802-a, 03-04-48829, 03-04-48555-a); RAS Presidium Program “Molecular and Cellular Biology (project No. 10.4); The Siberian Branch of the Russian Academy of Sciences (integration project No. 119); Russian Ministry of Industry, Science and Technologies (grants No. 43.073.1.1.1501).

References

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank // Nucleic Acids Res. 2000. V. 28. P. 235–242.

Bernstein F.C., Koetzle T.F., Williams G.J.B., Meyer E.F., Brice M.D., Rodgers J.R., Kennard O., Shimanouchi T., Tasumi M. The protein data bank: a computer-based archival file for macromolecular structures // J. Mol. Biol. 1977. V. 112. P. 535–542.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O’Donovan C., Phan I., Pilbout S., Schneider M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003 // Nucleic Acids Res. 2003. V. 31. P. 365–370.

Geutskens S.B., van den Wollenberg D.J.M., van der Eb M.M., van Ormondt H., Jochemsen A.G., Hoeben R.C. Characterisation of the p53 gene in the rat CC531 colon carcinoma // Gene Ther. Mol. Biol. 2000. V. 5. P. 81–86.

Greenblatt M., Bennett W., Hollstein M., Harris C. Mutations in the p53 tumor suppressor gene: clues to cancer etiology and molecular pathogenesis // Cancer Res. 1994. V. 54. P. 4855–4878.

Gualberto A., Aldape K., Kozakiewicz K., Tlsty T.D. An oncogenic form of p53 confers a dominant, gain-of-function phenotype that disrupts spindle checkpoint control // Proc. Natl Acad. Sci. 1998. V. 95. P. 5166–5171.

Hollstein M., Sidransky D., Vogelstein B., Harris C.C. p53 mutations in human cancers // Science. 1991. V. 253. P. 49–53.

Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A. PDBSite: a database on protein active sites and their environment // Third International Conference on Bioinformatics of Genome Regulation and Structure. 2002. V. 3. P. 145–148.

Ivanisenko V.A., Debelov V.A., Pintus S.S., Matsokin A.M., Nikolaev S.V., Grigorovich D.A., Kolchanov N.A. PDBSiteScan: A tool for search for the best-matching superposition in the database PDBSite // Third International Conference on Bioinformatics of Genome Regulation and Structure. 2002. V. 3. P. 149–152.

Iwanaga Y., Jeang K.-T. Expression of mitotic spindle checkpoint protein hsmad1 correlates with cellular proliferation and is activated by a gain-of-function p53 mutant // Cancer Res. 2002. V. 62. P. 2618–2624.

Kastan M.B., Onyekwere O., Sidransky D., Vogelstein B., Craig R.W. Participation of p53 protein in the cellular response to DNA damage // Cancer Res. 1991. V. 51. P. 6304–6311.

Levine A.J. The tumor suppressor genes // Annu. Rev. Biochem. 1993. V. 62. P. 623–651.

Lindahl E., Hess B., van der Spoel D. GROMACS 3.0: A package for molecular simulation and trajectory analysis // J. Mol. Mod. 2001. V. 7. P. 306–317.

Lowe S.W., Ruley H.E., Jacks T., Housman D.E. p53-dependent apoptosis modulates the cytotoxicity of anticancer agents // Cell. 1993. V. 74. P. 957–967.

Pennec X., Ayache N. A geometric algorithm to find small but highly similar 3d substructures in proteins // Bioinformatics. 1998. V. 14. P. 516–522.

Schwede T., Kopp J., Guex N., Peitsch M.C. SWISS-MODEL: an automated protein homology-modeling server // Nucleic Acids Res. 2003. V. 31. P. 3381–3385.

Sigal A., Rotter V. Oncogenic mutations of the p53 tumor suppressor: the demons of the guardian of the genome // Cancer Res. 2000. V. 60. P. 6788–6793.

Soussi T., Dehouch K., Beroud C. p53 website and analysis of p53 gene mutations in human cancer: forging a link between epidemiology and carcinogenesis // Hum. Mutat. 2000. V. 15. P. 105–113.

PERIODICAL PATTERNS IN SEQUENCES OF SPIDROINS I AND II AND SECONDARY STRUCTURE PREDICTION

Ragulina L.E. [1]*, ***Makeev V.Ju.***[2], ***Esipova N.G.***[3], ***Tumanyan V.G.***[3], ***Bogush V.G.***[2], ***Debabov V.G.***[2], ***Nikitin A.M.***[3], ***Vlasov P.K.***[3]

[1] Moscow Institute of Physics and Technology, Moscow, 141700, Russia; [2] State Scientific Center "GosNIIGenetika", Moscow, 113545, Russia; [3] Engelhardt Institute of Molecular Biology RAS, Moscow, 119991, Russia

* Corresponding author: e-mail: lera_846@pisem.net

Keywords: *Spider, web silk, protein, primary structure, periodical pattern, bioengineering, secondary structure prediction*

Summary

Motivation: Recombinant proteins of spider web are an important material for biotechnology. They have unique physicochemical properties. The valuable properties of a spider web fiber are likely to be related to the particular structure of protein they are made of. This makes important studying of the particular sequence structure of spider web proteins and their secondary structure.

Results: We have developed a technique of assessing patterns that are periodically distributed along the protein sequence. This method allowed us to identify the domain structure of spidroin sequences. We predicted the secondary structure of spidroins with three independent approaches. Our prediction supports the notion that protein is found in a gland in its globular state. The results of our prediction agree with the experimental results.

Availability: SymFour program is available from makeev@imb.ac.ru on request.

Introduction

Spider silk exhibits outstanding physical properties, making this fiber a promising raw material for different technological applications (1990). These physical properties are related to the particular structure of protein molecules the web fiber consists of (Madsen, 1999). The protein spatial structure at all organization levels is predefined by the sequence of its monomer units. Protein macro properties are associated with charge distribution along the polypeptide chain. In proteins, the charge distribution is associated with the amino acid distribution along the sequence, which in the fibrous proteins displays the periodic structure. This is not the case for the globular proteins. In this work we studied spidroin I and II proteins which make the backbone dragline of a spider web.

Methods and Algorithms

Fourier scanning: Program **catseq** was used to prepare sequence fragments of the protein sequence, in which the distribution of periodical patterns was studied with the help of symbolic Fourier transform program **symfour.** Sequence fragments were prepared sequentially from the N-terminus of the protein sequence. The length of the fragment, also called the length of the scanning window, was chosen to be greater than the total length of several characteristic segments of the spidroin protein sequence. We used overlapping scanning windows displaced one from another for the number of residues called the scanning step. The procedure of selecting sequence fragments was repeated until the remaining C-terminal fragment of the sequence became smaller than the scanning window. The results of symbolic Fourier transform computed by **symfour** were assembled in the table and graphically visualized.

The secondary structure of spidroin proteins was predicted from their amino acid sequence using three programs, which employ the PDB data bank as a training set: (i) NNPREDICT (www.cmpharm.ucsf.edu/~nomi/nnpredict.htm); (ii) JPRED (www.compbio.dundee.ac.uk/~www-jpred/) and (iii) OLIGON developed in Engelhardt Institute of Molecular Biology RAS.

Results

We analyzed all known long spidroin sequences obtained from spiders of different species with the help of the **symfour** program and have obtained the following results: the sequences of spidroins II include N-terminal, initial, middle, and C-terminal regions. For each region the particular periodical patterns of successive domains with insertion and deletions is characteristic. The only exception is the final part of the C-terminal region, which shows practically no structure. Dominant periods in the amino acid sequence are of 18–30 residues in N-terminal, initial and middle regions. In the C-terminal region dominate periods or 30–50 amino acid residues.

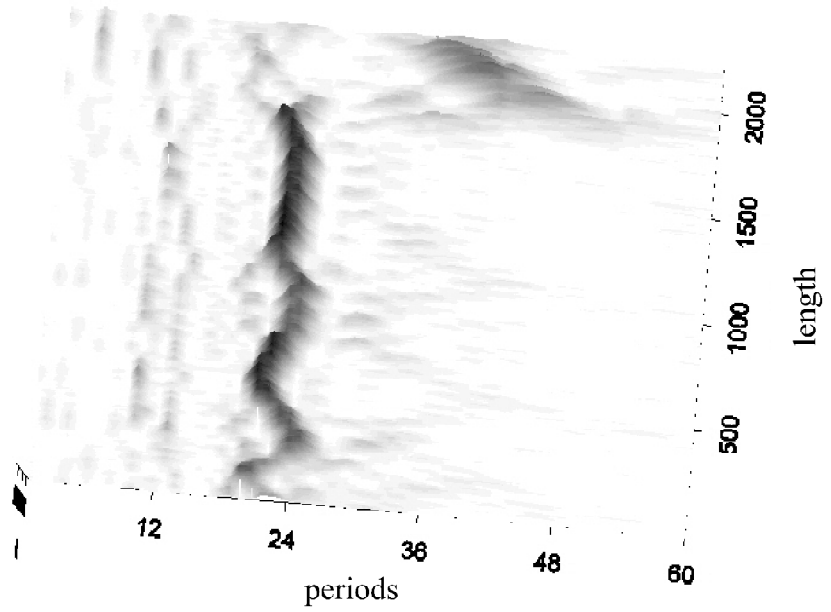


Fig. Periodical patterns of Spidroin 2 sequence of *N. madagascariensis.* The scanning window is 100 residues; the scanning step is one residue. The z-axis displays the spectral power, shown also with the shade. A darker shade corresponds to a higher spectral power.

The periodic diagrams were built for 6 pairs of sequences of different species spidroins (I and II). Secondary structure of spidroin I and II proteins was predicted with three independent methods.

1. The program **nnpredict** (www.cmpharm.ucsf.edu/~nomi/nnpredict.htm) yield a specific structure for 20 % of residues in this sequence. According to this prediction, amino acid residues L, E, Q and Y belonging to the characteristic oligopeptides QGYG, AGRGGLGA and RGEL adopt the local conformation of type β, whereas all A in poly-alanine blocs adopt α-helical conformation.

2. The JPRED server (www.compbio.dundee.ac.uk/~www-jpred/) yields results only for C-terminal spidroin sequence regions. The α-helix is predicted for some poly-alanine blocks,

whereas for a part of a conservative C-region sequence the following secondary structure is predicted:

CDVLVQALLEIVSALVHI
αααααααααααααααββ

3. Prediction made with the help of OLIGON developed in the EIMB RAS yields the following results: for spidroin I for poly-A blocks all alanines but the last one adopt the α-helical conformation, all prolines adopt the left helix of a polyproline II type. The protein contains the following basic oligopeptides with a characteristic secondary structure:

LGGQ, GAGQ, YGGL, RGGX	where X=Y, Q, L,
---β, -P--, P---, ---P,	P is the helix of polyproline II type;
	"–" denotes the absence of prediction.

The following combination of amino acid residues with a characteristic secondary structure were found for spidroin 2: QGPS, GYGP, PGQQ where S, P, Y, and Q adopt the conformation of a left helix of polyproline II type, GYGQ, where Y adopt the conformation of a β type, and PRQQ, where P and Q adopt the left helix conformation of a poly-proline II type and R adopts the α-helical conformation.

All three programs yield the same result only for the poly-A blocks and for the constant terminal sequence regions, which are conservative between all proteins. In spidroins I and II, in these regions there are some segments in which the secondary structure is predicted for all amino acids included:

CDVLVQALLELITALISI
Bβααβαααααααααααβ.

Discussion

Different sequence pattern in the Gly-rich domains in different chain segments is likely to be required to maintain the specific native aggregation of molecules in a fibril. Distinct periodic patterns characteristic for different segments result in the correct alignment of the protein molecules, instead of chaotic intermolecular interaction resulting in molecular mingling.

It is noteworthy that the programs performing secondary structure prediction are trained on globular proteins, whereas the spidroins are fibrous proteins. At the same time, recently there was reported that spidroins can form globules at the stage of secretion into the gland channel (Vollrath, Kinght, 2001). Our prediction of spidroin I and II secondary structure allows us to conclude that some molecule segments can adopt α-helix or left helix of polyproline II type, whereas the β-structure is rarer.

This observation agrees with the CD-studies for the engineered recombinant *Nephila clavipes* spidroin 1 analog.

Several authors reported that the β-helix is characteristic for the polyalanine regions (Simmons *et al*., 1996).

All in all, in spidroins we observe an interesting example of conformational transition from a globular structure to a fibrous structure. It is likely that protein aggregation into a fibril occurs simultaneously with this transition and the structure of a globular precursor is important for the emerging native structure of the fiber.

This knowledge can help to understand the arrangement of protein molecules in a fiber, which is required for synthesis of an artificial fiber with pre-given properties and for future obtaining of a spatial structure of a fiber.

Acknowledgements

This study is partially supported by the ISTF grant #1033, by RFBR grant 04-07-9027-2 and by Program in Molecular and Cellular Biology of Russian Academy of Sciences, coordinator V.G. Tumanyan.

References

Madsen B., Shao Z.Z., Vollrath F. Variability in the mechanical properties of spider silks on three levels: interspecific, intraspecific, and intraindividual // Int. J. of Biol. Macromol. 1999. V. 24. P. 301–306.

Makeev V.J., Tumanyan V.G. Search of periodicities in primary structure of biopolymers: a general Fourier approach // CABIOS. 1996. V. 12. P. 49–54.

Simmons A.H. *et al*. Molecular orientation and two – component nature of the crystalline fraction of spider dragline silk // Science. 1996. V. 271. P. 84–87.

Vollrath F., Kinght D.P. Liquid crystalline spinning of spider silk // Nature. 2001. V. 410. P. 541–548.

Xu M., Lewis R.V. Structure of a protein superfiber: spider dragline silk // Proc. Natl Acad. Sci. USA. 1990. V. 87(18). P. 7120–7124.

CONSTRUCTING DETAILED KNOWLEDGE-BASED ATOMIC POTENTIALS FOR WATER IN PROTEINS

Rahmanov S.V.*, Makeev V.Yu.

Laboratory of Bioinformatics, GosNIIGenetika, Moscow, Russia
***** Corresponding author: e-mail: sergeira@genetika.ru

Keywords: *protein, water, solute, hydration, atomic potentials, knowledge-based*

Summary

Motivation: Proper estimate of protein solvation energy is a crucial factor for protein folding modeling and function analysis. Novel continuous knowledge-based potentials for water to atom interaction were developed, based on statistics of pairwise distances between water molecules reported in PDB structures and protein atoms of different types.

Results: A large training set of 1218 protein structures with sequence identity under 25 % was used, and a new method was elaborated for creating comparable reference state for each structure. This allowed us to obtain continuous, highly discriminating atomic potentials, which behave reasonably for hydrophilic and hydrophobic residues. The potentials were successfully tested by calculating estimated water energies at actual water positions as compared to random locations, and also by comparing calculated hydration energies for the whole native proteins with those for the misfolded structures.

Availability: http://bioinform.genetika.ru/projects/projects_en.html

Introduction

Solvent plays an essential role in protein folding and function. A number of methods were developed for modeling solvent and solute-solvent interactions including methods based on solvent-accessible surface area (Richmond, 1984), and the hydration shell volume (Kang *et al*., 1988). How well these methods can mimic the solvent environment and its influence is still unclear.

We developed a method that uses information on water molecules reported in the 3D structures of proteins, the so-called "bond" or "conserved" water. X-ray crystallographic studies provide only the coordinates of the water oxygen atom. The large number of water molecules reported in the PDB database (there is about one water molecule reported per two amino acid residues) offers a sufficient amount of statistical data to construct detailed pairwise distance distributions between water molecules and atoms of different types in the protein structure. The atom type is defined by its residue, name, and position within the residue; for instance, the CA (alpha-carbon) atom of leucine, CA of asparagin and CB (beta-carbon) of leucine are considered as separate types. The number of water molecules per amino acid differed significantly (in some cases, for more than 20 times) among protein structures reported. To reduce the influence of the inhomogeneity of the database we developed a method for normalizing the data obtained from different structures, calculating the distance distribution from water molecules reported for the structure to random points within 6.5 Å first hydration shell. This reference distribution was calculated by the Monte Carlo method with the number of test points sufficient to obtain a saturated distribution. The distribution shape is specific for each individual structure and determined by its geometry and water content. This approach also solves the key problem in the derivation of knowledge-based potentials: choosing the appropriate reference state (Lu, Skolnick, 2001). To measure if the water molecule is preferred in the vicinity of an atom of a certain type, for each protein structure we calculated the empirical distance distribution between all such atoms and all water molecules in

this structure, normalized it for the reference distribution, and combined the normalized distributions for all protein structures in the database. The resulting distance distribution thus becomes independent from the geometry of the protein structure, amino acid composition, and water content of an individual structure and depends only on the atom type. In the quasi-chemical approach we utilize, a distributions of distances from an atom of a particular type to water molecules is used to evaluate the atomic hydration potential. The potentials obtained in this study can be used to predict locations of molecules of bound water in proteins. Another value, which can be calculated using these potentials, is the share of time for which a particular vicinity of a protein structure is occupied by at least one water molecule. The potentials for different atom types can be added resulting in a cumulative hydration potential for the whole structure, allowing estimation of the contribution of the hydration energy into different protein folding variants.

Methods and Algorithms

At distance d the hydration potential is calculated with the formula:

$$\varepsilon(d) = -RT \ln\left(\frac{D_{obs}}{D_{\exp}}\right), \quad (1)$$

where D_{obs} is the observed spatial density of water at distance d, and D_{exp} is the reference density. In the hypothetical case of the training set consisting of one protein, for every atom type a, the atomic potential at distance d is given by the formula:

$$\varepsilon_a(d) = -RT \ln \sum_{i=1}^{n} \frac{D_a(i)(d)/n}{D_P^r(d)}, \quad (2)$$

where n is the number of atoms of type a, $D_a(d)$ is the estimated spatial density of water at distance d for each atom, and D_p^r is the estimated reference distribution. After normalizing for the reference distribution the resulting atom type specific distribution takes the form:

$$D(a) = \sum_{P} \sum_{i=1}^{n} \frac{D_a(i)/n}{D_P^r}. \quad (3)$$

This formula indicates how the presence of an atom of this type effects the probability of the water molecule occurrence at a particular distance from the atom.

Implementation and Results

A subset of PDB, 1218 proteins with no chain pair of sequence identity greater than 25 % (Hobohm, Sander, 1994) and with water content of no less than one water molecule per 30 residues, was used as the training set for developing knowledge-based hydration potentials. Crystallographic symmetry-related structures were generated for better representation of atomic hydration environment before obtaining the pairwise distance data. For each protein, the estimated distance distribution from random location to water was obtained in a direct modeling experiment, averaging over a large number of random probes. From the 8936 distinct atom types found in the training set, 342 most represented protein atom types and 29 hetero-atoms such as metal atoms (i.e., Zn) or inorganic groups atoms (SO4 and so on) were selected. The number of pairwise distances used to create the potential distributions ranged from hundreds of thousands to tens of millions for different atom types.

Four normalized distributions for NE2 atom of glutamine, CD1 of isoleucine, HE1 atoms of histidine and phenylalanine, are shown in Figure 1. While the first two "hydrophilic" atoms have a characteristic peak at close distances and a decline at a medium range, the latter two atoms exhibit a peak at a middle distance range typical for hydrophobic residue atoms.

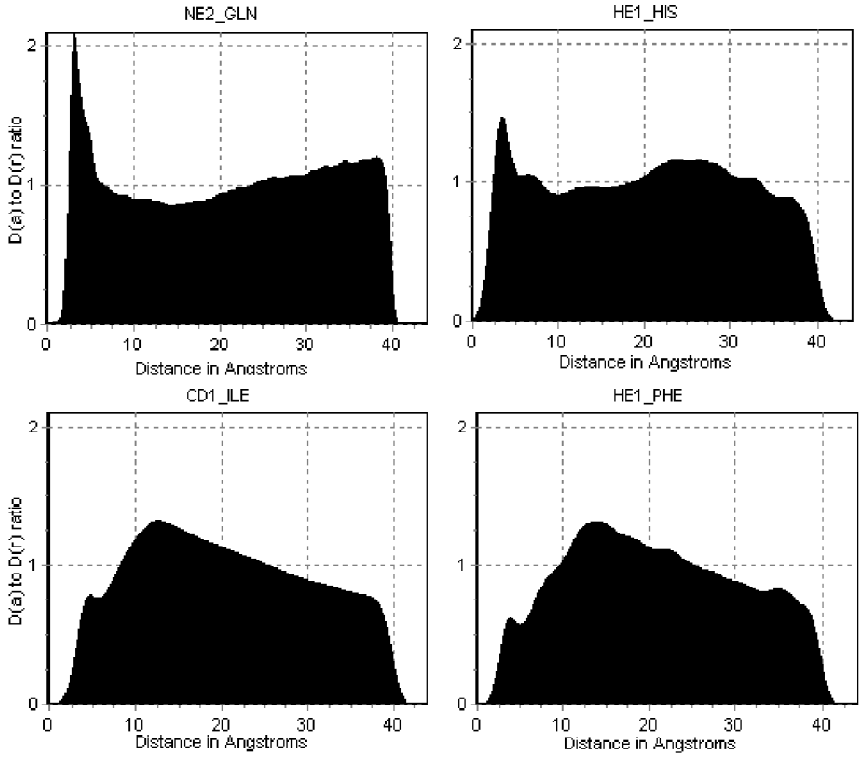


Fig. 1. The normalized radial distribution density of water around the atom ($D(a)$). The atom type is indicated above each diagram. Two upper diagrams show distributions for atoms that are partially charged and "hydrophilic", while the two lower diagrams show distributions for neutral atoms from "hydrophobic" residues.

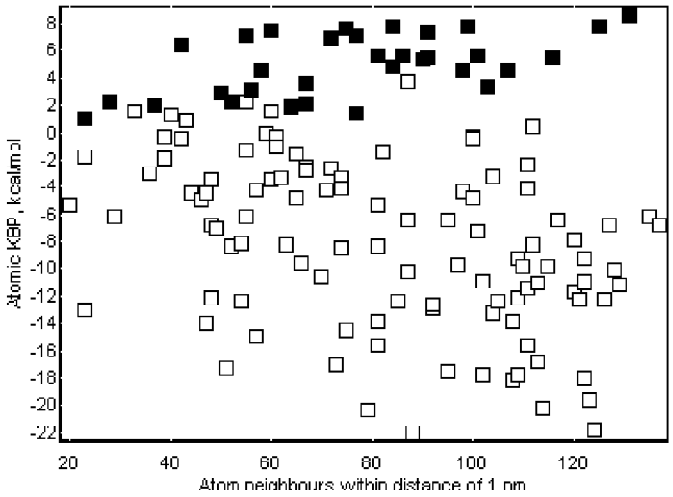


Fig. 2. The comparison of hydration energies calculated for actual water positions in a typical protein (filled rectangles) with that calculated for random locations (blank rectangles). The number of neighbor atoms included in the calculation is given along the horizontal axis. KBHP is the value of the knowledge-based hydration potential.

To test the results we compared the hydration energies calculated for the actual positions of water molecules in proteins not included into the training set, with those calculated for random locations with similar space constraints (no atoms in the close vicinity etc). The typical example is shown in Figure 2.

The atomic potentials developed in this study have also been tested on the protein decoy sets posted on the PROSTAR website, http://prostar.carb.nist.gov. The objective of this test is to discriminate between the native structure and one or more misfolded decoy structures. The atomic potentials presented proved to consistently favor the native folds over decoys.

Discussion

The potentials developed can be viewed as a useful tool for solving several problems like predicting water location sites and propensities, quantitative assessment and visualization of solvent-accessible regions of molecular structure, and as a tool for estimating the contribution of hydration energy in protein folding modeling.

References

Hobohm U., Sander C. Enlarged representative set of protein structures // Protein Science. 1994. V. 3. P. 522.

Hui Lu., Jeffrey Skolnick. A Distance-Dependent Atomic Knowledge-Based Potential for Improved Protein Structure Selection // PROTEINS: Structure, Function, and Genetics. 2001. V. 44. P. 223–232.

Kang Y.K., Gibson K.D., Nemethy G., Scheraga H.A. Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell model // J. Phys. Chem. 1988. V. 91. P. 4739–4742.

Richmond T.J. Solvent accessible surface area and excluded volume in proteins, analytical equations for overlapping spheres and implications for the hydrophobic effect // J. Mol. Biol. 1984. V. 178. P. 63–89.

MINING FROM COMPLETE PROTEOMES TO IDENTIFY ADHESINS AND ADHESIN-LIKE PROTEINS: A RAPID AID TO EXPERIMENTAL RESEARCHERS

Ramachandran S.*, Jain P., Sachdeva G.

G.N. Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India

***** Corresponding author: e-mail: Ramu@igib.res.in; ramucbt@yahoo.com

Keywords: *proteomes, adhesins, neural networks, composition, pathogens, microbes, SARS, bacteria, non-homology, enteropathogenic E. coli*

Summary

Motivation: Adhesion of a microbial pathogen to a host is mediated by adhesins. Currently, experimental methods are used for detecting and characterizing adhesins, which are time consuming and demand large resources. The availability of a software specifically focused to identify adhesins can aid in rapid experimentation. We have employed artificial neural networks to develop SPAAN, which predicts the probability of a protein being an adhesin (Pad) based on 105 compositional properties of its sequence.

Results: SPAAN could identify 97.4 % of known adhesins at high Pad value from many bacteria and also guided in improving the annotation of several proteins as adhesins. Several novel adhesins were identified in 17 pathogenic organisms. In the Severe Acute Respiratory Syndrome (SARS) associated human corona virus, the spike glycoprotein and nsp's (nsp1, nsp5, nsp6 and nsp7) were identified with adhesin-like characteristics. These results offer new leads for rapid experimental testing.

Availability: SPAAN is freely available for academic users through ftp from either 203.195.151.45 or 203.90.127.75. user: ramu; password: ramu897; ramu@igib.res.in

Introduction

Microbial pathogens encode several proteins known as adhesins that mediate their adherence to host cell surface receptors, membranes, or extracellular matrix for successful colonization. New approaches to vaccine development focus on targeting adhesins to abrogate the colonization process (Wizemann *et al.*, 1999). Examples are the currently approved vaccine against whooping cough caused by *B. pertussis* contains filamentous hemagglutinin and pertactin adhesins (Halperin *et al*., 2003). Immunization with FimH is being evaluated to protect against pathogenic *E. coli* (Langermann *et al*., 2000). Likewise, outer membrane vesicle preparations are being tested against gastric ulcer caused by *H. pylori* (Prinz *et al*., 2003).

Prediction of adhesins or adhesin-like proteins is important for the development of new vaccine formulations. Experimental identification of adhesins is an arduous task. Computer based search softwares can accelerate the prediction of adhesins. We report SPAAN, a non-homology method using artificial neural networks (ANN) to identify adhesins and adhesin-like proteins in diverse species.

Methods and Algorithms

1. Amino acid frequency f_i = (counts of i^{th} amino acid) / l; i = 1…20; l is the length of the protein.

2. Multiplets were identified according to Brendel *et al*. (1992), frequencies of the amino acids in the multiplets $f_i(m)$ = (counts of i^{th} amino acid in multiplet) / l.

3. Dipeptide (i, j) frequency f_{ij} = (counts of ij^{th} dipeptide) / (total dipeptide counts); i, j ranges from 1 to 20. Top 20 dipeptides that served to discriminate adhesins from non-adhesins (assessed by *t-test*) were used: NG, RE, TN, NT, GT, TT, DE, ER, RR, RK, RI, AT, TS, IV, SG, GS, TG, GN, VI, and HR.

4. Charged amino acids (R, K, E and D at pH 7.2) frequency, f_c = (counts of charged amino acids) / l. Distribution of the charged amino acids were estimated using moments of the positions of the occurrences these amino acids. Moments of order 'r' is $M_r = \sum (X_i - X_m)^r/N$. X_m = mean of all positions of charged amino acids; X_i = position of i^{th} charged amino acid; N = number of charged amino acids. The 20 inputs were: f_c, l, M_r (r = 2 to 19).

5. Hydrophobic composition: Amino acids grouped into five classes by (Brendel *et al*., 1992): (-8 for K, E, D,R), (-4 for S, T, N, Q), (-2 for P, H), (+1 for A, G, Y, C, W), (+2 for L, V, I, F, M).

A total of 25 inputs were fed to ANN. (a) f_i = (counts of i^{th} group) / (total counts in the protein); i ranges from 1 to 5. (b) m_{ji} = j^{th} order moment of positions of amino acids in i^{th} group; j ranges from 2 to 5. A grand total of 105 compositional properties were used.

For training the network, protein sequences were retrieved from http://www.ncbi.nlm.nih.gov using the keyword 'adhesin'. The records from primary retrieval were thoroughly curated through manual examination to produce a set of well-annotated adhesins. For non-adhesins, we used the rationale to collect sequences of enzymes and other proteins that function within the cell. Redundant entries were examined using CLUSTALW (Thompson, 1994) and were eliminated.

The neural network used here has a multi-layer feed forward topology. It consists of an input layer, a hidden layer and an output layer. The feed forward error back propagation neural network algorithm was used. The program was downloaded from the web site (http://www.cs.colostate.edu/~anderson) a kind gift from Charles W. Anderson, Department of Computer Science, Colorado State University, Fort Collins, CO 80523, anderson@cs.colostate.edu.

A network was trained optimally for each attribute. Five networks were prepared. Query proteins were processed modularly through five networks. The final probability (P_{ad}) is $P_{ad} = (P_A * fc_A + P_C * f_{cc} + P_D * fc_D + P_H * fc_H + P_M * fc_M) / (fc_A + fc_c + fc_D + fc_H + fc_M)$ fc_i = fraction of correlation of i^{th} network; i = A (Amino acid frequencies), C (Charge composition), D (Dipeptide frequencies), H (Hydrophobic composition), or M (Multiplet frequencies), $fc_A = 0.84$, $fc_C = 0.71$, $fc_D = 0.84$, $fc_H = 0.79$, $fc_M = 0.83$. Computer programs were written in C and executed on a PC with Red Hat Linux ver 7.3 or 8.0 operating system.

Sequences should be in FASTA format and in Single letter code (IUPAC-IUB nomenclature) either singly or multiply. Protein sequences with ambiguous amino acids (other than 20 amino acids) and/or of length less than 50 amino acids are filtered out.

Implementation and Results

Sensitivity (*Sn*), *Specificity (Sp) and Correlation coefficient* (*CC*)

SPAAN is a non-homology, compositional property based method to predict adhesins and adhesin-like proteins solely from sequence data. It had sensitivity 89 % and specificity 100 % at P_{ad} 0.51. At this threshold P_{ad}, the Matthew's correlation coefficient (Matthews, 1975) was observed to be highest (0.94).

SPAAN predicts experimentally characterized adhesins with high P_{ad} value

Of 194 known adhesins from several bacterial pathogens 189 adhesins had P_{ad} i 0.51, indicating an overall sensitivity of 97.4 %. These results demonstrate the general applicability of SPAAN to detect a wide range of adhesins from diverse bacteria.

Application of SPAAN to whole genomes

The results of the genome scan of a representative human pathogen is displayed in Table. Detailed

manuscript has been submitted (Sachdeva *et al*., 2004). We used stringent criterion of $P_{ad} > 0.7$ to minimize the detection of false positives. Further, the top 50 proteins identified by SPAAN were prioritized and analyzed thoroughly. Several of the predicted adhesins are supported by complementary evidence using Conserved Domain Database search (RPS-BLASTP), BLASTP and beta-wrap (Marchler-Bauer *et al*., 2002; Altschul *et al*., 1990; Bradley *et al*., 2001). The top scoring proteins in *E. coli* O157:H7 include 9 Fimbrial adhesins, AidA-I, Gamma intimin, Hemagglutinin, translocated intimin receptor all which have been characterized. In addition several novel adhesins were predicted. The results on several other pathogens have been described in detail in the submitted manuscript (Sachdeva *et al*., 2004).

Table. Analysis of predictions made by SPAAN on complete proteome of *E. coli* O157:H7. (Results on other pathogens are available from Sachdeva *et al*., 2004, submitted)

Organism	Disease caused	Total No. of Proteins Analyzed	No. of these Supported by Complementary Evidence CDD/ blastp/PubMed	No. of these Supported by Complementary evidence BetaWrap	No. of adhesion like proteins	No. of false positives
E. coli O157:H7	Diarrhoea	50	37*	33	12	1

* Includes Fimbrial adhesins [9 proteins], AidA-I, Gamma intimin, Hemagglutinin, translocated intimin receptor, putative tail fiber protein, putative major tail protein.

Discussion

At present there is no software available for identifying adhesins in complete proteomes. The usual method employed is a combination of homology search and perhaps the RPS-BLASTP. We have attempted to address this important issue and have developed an algorithm, which is based on 105 compositional attributes of proteins (Sachdeva *et al*., 2004 submitted). The algorithm described in this work called SPAAN is a non-homology based method. The algorithm not only identified known adhesins but also guided in improved annotation of several proteins and aided in the identification of novel adhesins.

An example of a pathogenic *E. coli* complete proteome processed through SPAAN has been displayed in Table. It is apparent that the results are very encouraging with very low false positives. We have scanned over 20 proteomes by now but restricted space does not permit us to discuss results from numerous organisms. Compositional properties have now been used to identify the sub-cellular location of proteins, the secreted proteins, apicoplast targeted proteins, functional class identification, and therefore is emerging as a general theme for tackling cases that met with defeat while using homology based methods.

Acknowledgements

We thank Charles Anderson for Neural Network program. SR thanks the Council of Scientific and Industrial Research (CSIR) for grants under (NMITLI).

References

Altschul S.F. *et al*. Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. P. 403–410.

Bradley P., Cowen L., Menke M., King J., Berger B. BETAWRAP: successful prediction of parallel beta - helices from primary sequence reveals an association with many microbial pathogens // Proc. Natl Acad. Sci. USA. 2001. V. 98. P. 14819–14824.

Brendel V., Bucher P., Nourbakhsh I.R., Edwin Blaisdell B., Karlin S. Methods and algorithms for statistical analysis of protein sequences // Proc. Natl Acad. Sci. USA. 1992. V. 89. P. 2002–2006.

Halperin S.A. *et al*. Nature, evolution, and appraisal of adverse events and antibody response associated with the fifth consecutive dose of a five-component acellular pertussis-based combination vaccine //

Vaccine. 2003. V. 21. P. 2298–2306.

Langermann S. *et al*. Vaccination with FimH adhesin protects cynomolgus monkeys from colonization and infection by uropathogenic *Escherichia coli* // J. Infect. Dis. 2000. V. 181. P. 774–778.

Marchler-Bauer A. *et al*. CDD: a database of conserved domain alignments with links to domain three-dimensional structure // Nucleic Acids Res. 2002. V. 30. P. 281–283.

Prinz C., Hafsi N. Voland P. *Helicobacter pylori* virulence factors and the host immune response: implications for therapeutic vaccination // Trends in Microbiol. 2003. V. 11. P. 134–138.

Sachdeva G., Kumar K., Jain P., Ramachandran S. SPAAN: A Software for Prediction of Adhesins and Adhesin-like proteins using Neural networks. Manuscript submitted to Bioinformatics. 2004.

Thompson J.D., Higgins D.G., Gibson T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. V. 22. P. 4673–4680.

Wizemann T.M. *et al*. Adhesins as targets for vaccine development // Emerg. Infect. Dis. 1999. V. 5. P. 395–403.

MUTANT PROTEIN STRUCTURES REVEAL MOLECULAR MECHANISMS OF INHERITED DISEASES

Ramensky V.E.*, Tumanyan V.G.

Engelhardt Institute of Molecular Biology of RAS, Moscow 119991 Vavilova, 32, Russia
***** Corresponding author: e-mail: ramensky@imb.ac.ru

Keywords: *human disease, protein structure, protein function, amino acid substitution, local deformation of protein structure, disease mutation*

Summary

Motivation: Comparative analysis of pairs of protein structures for which phenotypes resulting from mutation are known can help provide interpretation of molecular mechanisms underlying inherited traits.

Results: The data derived from comparison of mutant and wild-type protein structures suggest that local deformation of protein structure around amino acid replacement site is one of phenomena observed in mutant protein forms associated with human diseases. Understanding mechanisms of its formation can help unravel the molecular basis of inherited diseases.

Introduction

Interpretation of effects of a mutation in a protein is a non-trivial task usually based upon a combination of sequence and structure-derived characteristics (Sunyaev *et al*., 2001; Chasman, Adams, 2001; Ramensky *et al*., 2003). The proper functioning of a protein requires that its critical structural features are not affected. This makes analysis of three-dimensional structures of proteins containing mutations a powerful tool for exploration of molecular basis of inherited human diseases (Wang, Moult, 2001; Steward *et al*., 2003). This analysis, however, is very often confined to single structures representing only wild-type or mutant protein. In this work we present a survey of pairs of protein structures for which phenotypes resulting from mutations are known. This enables direct study of differences between wild-type and mutant structures and interpretation of molecular mechanisms underlying inherited traits. In particular, we show that local deformation of protein structure around amino acid replacement site is one of phenomena observed in mutant protein forms associated with human diseases.

Methods and Algorithms

The data on mutations in proteins were extracted from literature-based Protein Mutant Database (Kawabata *et al*., 1999) which is a compilation from approximately 15,000 articles published since 1970s. The database currently contains descriptions of about 95,000 unique mutations in approximately 6,000 proteins. The protein tertiary structures were taken from Protein Data Bank (Berman *et al*., 2000) that contains 13,838 unique protein sequences. We performed BLAST search (Altschul *et al*., 1990) of PMD against PDB with the conditions that (a) both wild-type and mutant sequence have an exact match to sequences of different structures from PDB and (b) the corresponding phenotype is described in PMD. The structural parameters of mutant sites were determined with help of DSSP database (Kabsch, Sander, 1983) and ProCheck software (Laskowski *et al*., 1993). The regions around mutation site were structurally superimposed with a routine that implements the standard RMSD-minimizing orthogonal transform (Shapiro *et al*., 1992).

Implementation and Results

The described procedure results in a set of structure pairs representing 1,157 mutations in 333 proteins; of these, 888 pairs correspond to single point mutations. The most frequent phenotype

description fields in this set are STRUCTURE (912 occurences), FUNCTION (829), STABILITY (471), and EXPRESSION (43). Association with disease (annotated as DISEASE) is clearly stated only for 23 point mutations in 11 human proteins. The data for these cases are summarized in Table. The spatial superposition of regions flanking the substitution site was performed for structure pairs from Table. In 9 cases of 23, the distributions of pairwise distances between corresponding Ca-atoms in the flanking regions suggest that we observe the local deformation of structure. These cases are marked with asterisk in Table.

Table. Disease-associated mutations in human proteins for which both the wild-type and mutant protein structures were found. Figures in the "Mutation" column correspond to sequence numbering. Asterisks mark substitutions for which local structural deformation is observed

No.	Protein	Mutation	PDB id wild	PDB id mutant	Disease
1	Serum albumin	R242H	1n5uA	1hk2A	Familial dysalbuminemic hyperthyroxinemia
2		R242P		1hk3A	
3	Apolipoprotein E3	K164E	1nfn_	1ea8A	Type III hyperlipoproteinamia
4		K164Q		1h7iA	
5		R176C *		1le2_	
6	Breast cancer associated BRCA1	M1775R *	1jnxX	1n5oX	Breast/ovarian cancer
7	Tyrosine-protein kinase BTK	E41K	1b55A	1bwnA	Immunodeficient phenotype
8	Fibroblast growth factor 2	S252W *	1ev2E	1ii4E	Apert syndrome
9	Lysozyme	D85H	1c46A	1lyy_	Amyloidosis
10		I74T		1oua_	
11	Prion protein	E200K	1qlzA	1fo7A	Creutzfeldt-Jakob disease
12	Transforming protein P21/H-RAS-1	G12V *	1iozA	2q21_	Primary squamous cell carcinomas
13	Cu/Zn superoxide dismutase	G37R	1hl5A	1azvA	Familial amyotrophic lateral sclerosis
14		H46R *		1oezW	
15	Transphyretin	A129T	1f41A	1etb1	Familial amyloidotic polyneuropathy
16		L75P *		5ttrA	
17		S97Y *		2tryA	
18		T80A		1tshA	
19		V142I *		1ttrA	
20		V50M		1ttcA	
21		Y134C *		1iiiA	
22	Von Willebrand factor	I1309V	1oakA	1ijbA	Unusual type 2B phenotype of van Willebrand disease
23		R1306Q		1m10A	

Discussion

The survey of deleterious mutations and non-synoymous SNPs shows that there exist many different mechanisms by which a mutation can affect a structural feature (Wang, Moult, 2001; Steward *et al*., 2003; Ramensky *et al*., 2003). Our dataset suggests that local deformation of a protein structure around amino acid replacement site is one of the major mechanisms affecting phenotype. Indeed, the Leu55Pro mutation in human transphyretin eventually results in formation of amyloid fibrils (Sebastiao *et al*.,1998) and is the most pathogenic form among more than 70 others observed in familial amyloidotic polyneuropathy (Hornberg *et al*., 2000). This mutation is a straighforward

example showing that local deformation of protein tertiary structure may disrupt quaternary structure formation. In human lysozyme, Asp85His variant produces local structural deformation and amyloid fibrillogenesis (Booth *et al*., 1997). In the case of human Cu/Zn superoxide dismutase, the His46Arg mutant plays a key role in progression of familial amyotrophic lateral sclerosis (Deng *et al*., 1993). The observed structural deformation probably disrupts the proper positioning of Cu and Zn ions. As a result, the mutant form almost completely loses its function (<1 % of the wild-type activity). These examples suggest that unraveling the mechanism of formation of local structural deformation can help understand the molecular basis of inherited diseases.

Acknowledgements

This work was supported by "Molecular and Cell Biology" Grant Programme from Presidium of RAS and the Grant No. 43.071.1.1.1517 from Ministry of Industry, Science and Technology of Russian Federation.

References

Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. Basic local alignment search tool // J. Mol. Biol. 1990. V. 215. P. 403–410.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The protein data bank // Nucleic Acids Res. 2000. V. 28(1). P. 235–42.

Booth D.R., Sunde M., Bellotti V., Robinson C.V., Hutchinson W.L., Fraser P.E., Hawkins P.N., Dobson C.M., Radford S.E., Blake C.C., Pepys M.B. Instability, unfolding and aggregation of human lysozyme variants underlying amyloid fibrillogenesis // Nature. 1997. V. 385. P. 787–793.

Chasman D., Adams R.M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation // J. Mol. Biol. 2001. V. 307. P. 683–706.

Deng H.X. *et al*. Amyotrophic lateral sclerosis and structural defects in Cu,Zn superoxide dismutase // Science. 1993. V. 261. P. 1047–1051.

Hornberg A., Eneqvist T., Olofsson A., Lundgren E., Sauer-Eriksson A.E. A comparative analysis of 23 structures of the amyloidogenic protein transthyretin. 2000.

Kabsch W., Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. 1983. V. 22. P. 2577–2637.

Kawabata T., Ota M., Nishikawa K. The protein mutant database // Nucleic Acids Res. 1999. V. 27. P. 355–357.

Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M. PROCHECK: a program to check the stereochemical quality of protein structures // J. Appl. Cryst. 1993. V. 26. P. 283–291.

Ramensky V., Bork P., Sunyaev S. Human non-synonymous SNPs: server and survey // Nucleic Acids Res. 2003. V. 30. P. 3894–3900.

Sebastiao M.P., Saraiva M.J., Damas A.M. The crystal structure of amyloidogenic Leu55 —> Pro transthyretin variant reveals a possible pathway for transthyretin polymerization into amyloid fibrils // J. Biol. Chem. 1998. V. 273. P. 24715–2422.

Shapiro A., Botha J.D., Pastore A., Lesk A.M. A method for multiple superposition of structures // Acta Cryst. 1992. A48. P. 11–14.

Steward R.E., MacArthur M.W., Laskowski R.A., Thornton J.M. Molecular basis of inherited diseases: a structural perspective // Trends Genet. 2003. V. 19. P. 505–513.

Sunyaev S., Ramensky V., Koch I., Lathe W. 3rd, Kondrashov A.S., Bork P. Prediction of deleterious human alleles // Hum. Mol. Genet. 2001. V. 10. P. 591–597.

Wang Z., Moult J. SNPs, protein structure, and disease // Hum. Mutat. 2001. V. 17. P. 263–270.

BENCHMARKING OF TRANSMEMBRANE HELIX PREDICTION SERVERS

Sadovskaya N.S.

GosNIIGenetika, Moscow, Russia, e-mail: natasha@imb.imb.ac.ru

Keywords: *bacteria, transmembrane helix, prediction, benchmarking, cluster*

Introduction

Transmembrane helices (TM helices) in integral membrane proteins consist of regions of 15–30 predominantly hydrophobic residues which form б-helices separated by polar linking chains. A transmembrane protein may have one TM helix (membrane anchor with which the protein attaches to the membrane) or several TM helices that often form a membrane channel (transporters, many receptors, reductases and pumps). Several algorithms exist for identification of potential TM helices. These methods achieve a 90–95 % true positive rate with a false positive rate of only a few percents (von Heijne, 1996*)*. We benchmark ten most widely used TM-helix prediction servers using the consistency criterion: predictions for homologous proteins should be similar.

Methods

TM helix prediction algorithms use the following features: a) TM helices are predominantly non-polar; b) their length varies from 12 to 35 residues; c) the natural topology of a TM domain is helix – internal loop – helix – external loop – helix – internal loop, etc.; d) globular regions between membrane helices are usually shorter than 60 residues; e) internal non-membrane regions are "more positively charged" than the external nonmembrane regions, the "positive external rule" (von Heijne, 1986).

For benchmarking of TM servers we make the following assumptions: a) the number and lengths of TM segments is conserved in a group of orthologous proteins; b) when these proteins are aligned, the positions of the TM helices also should coincide. That is, corresponding TM helix in all proteins should map to one segment of the alignment (during alignments the relative positions of TM helices are conserved).

There are many servers that predict TM helices in protein sequences:

DAS http://www.sbc.su.se/~miklos/DAS/;
SPLIT http://garlic.mefos.hr/split/;
PRED-TMR http://o2.biol.uoa.gr/PRED-TMR/;
SOSUI http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html;
TMAP http://www.mbb.ki.se/tmap/single.html;
TopPred II http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html;
PSORT http://psort.ims.u-tokyo.ac.jp/form.html;
SAPS http://www.isrec.isb-sib.ch/software/SAPS_form.html;
TMPred http://www.ch.embnet.org/software/TMPRED_form.html;
HMMTOP http://www.enzim.hu/hmmtop/index.html;
TMHMM2.0 http://www.cbs.dtu.dk/services/TMHMM-2.0/;
MEMSAT http://www.sacs.ucsf.edu/secure/cgi-bin/memsat.pl;
Psi-Pred http://bioinf.cs.ucl.ac.uk/psipred/psiform.html;
PHDhtm http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_htm.html;
OrienTM http://biophysics.biol.uoa.gr/orienTM/;
ConPred http://bioinfo.si.hirosaki-u.ac.jp/~ConPred/;
CoPreTHi http://biophysics.biol.uoa.gr/CoPreTHi/.

In this work we analyzed 10 of these: DAS, PRED-TMR, SOSUI, TMAP, TopPred II, PSORT, TMPred, HMMTOP, TMHMM2.0, PHDhtm. The selection criteria were: a) server availability; b) the ability to process many requests; c) the ability to predict TM helices in a single protein rather than for a multiple alignment.

For comparison of several TM helices in two proteins it is useful to define two indices. The first index, QQ, compares TM segments on the amino acid level (i.e. it is a continuous index). For a pair of aligned TM-proteins, the QQ should be close to 1. The second index, KFS, characterizes the number of overlapping TM helices in compared proteins. Again, *KFS* =1 for absolutely consistent predictions. Overall, 3265 pairs of proteins were analyzed.

The obtained sequences were aligned with ClustalW (Thompson *et al*., 1994) and the positions of TM helices were predicted by servers listed above. For all methods we used the default parameters.

The TC.2A bacterial transporter dataset. For our analysis we collected all members of TC.2A class (Saier, 1999) and their homologs in sufficiently completed genomes. The dataset was divided into clusters (Sutormin, 2003) using the nearest neighbor clustering procedure based on the BLAST identity level. We considered clusters with the identity level of 40–49 % and the identity level of 50–59 %. Clusters with an identity level less than 40 % and less than 59 % were not considered.

Results and Discussion

Members of each group of orthologous genes were compared pairwise. Overlapping or adjacent TM helices predicted by a single server were considered as a single helix. For each server we computed QQ and KFS values of the predicted TM helices.

The best results were demonstrated the PHDhtm server. The predictions of TopPred, TMHMM2.0 and HMMTOP are of comparable high consistency. SOSUI, Pred-TMR and TMPred form the "middle layer" between the relatively better and worse servers. The servers Psort, DAS and TMAP showed the least consistent results. Figures 1 and 2 show the average values of QQ and KFS for the proteins in each cluster.

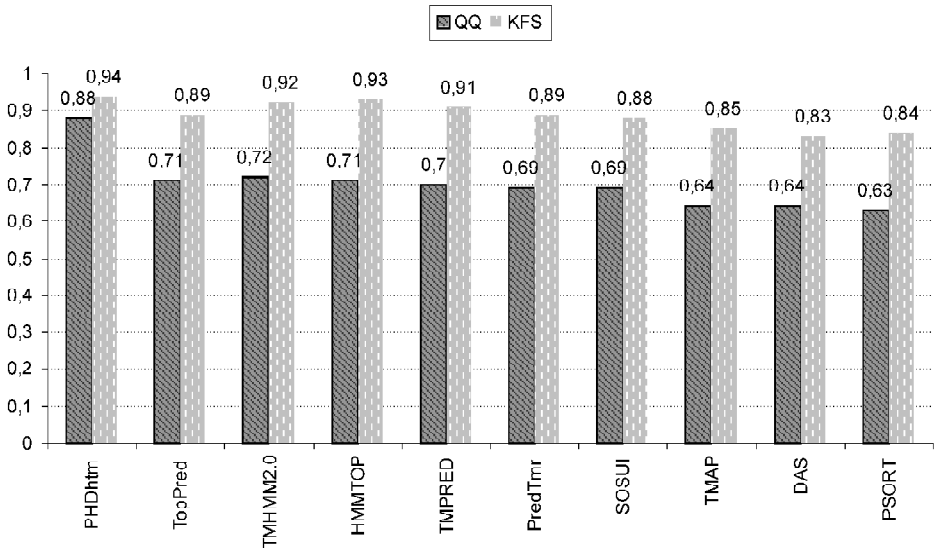


Fig. 1. The self-consistency of predictions of the servers DAS, HMMTOP, PHDhtm, PRED-TMR, PSORT, SOSUI, SPLIT, TMAP, TMHMM2.0, TMPred and TopPred in clusters corresponding to the identity level 40–49 %. The average QQ and KFS values are shown.

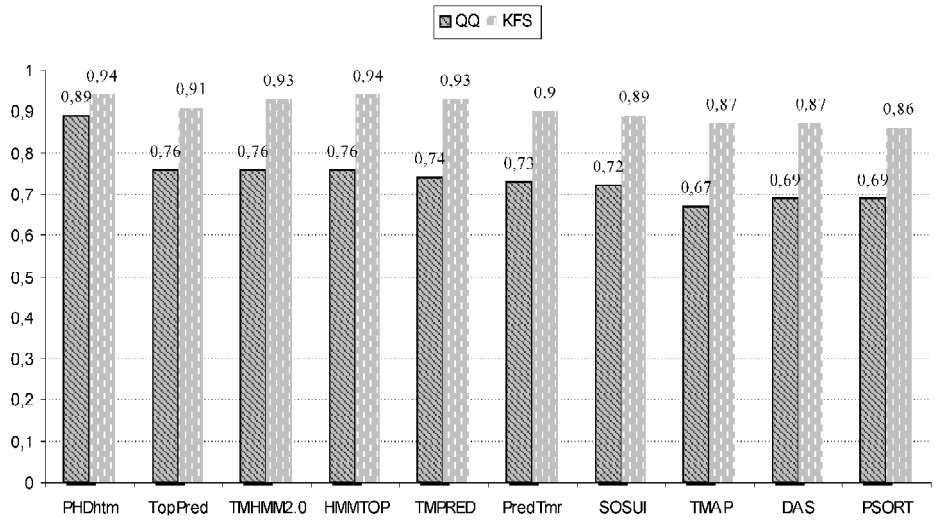


Fig. 2. The self-consistency of predictions of the servers DAS, HMMTOP, PHDhtm, PRED-TMR, PSORT, SOSUI, SPLIT, TMAP, TMHMM2.0, TMPred and TopPred in clusters corresponding to the identity level 50–59%. The average QQ and KFS values are shown.

Acknowledgements

This is joint work with M.S. Gelfand. We are grateful to R.A. Sutormin for the help with the data and to A.A. Mironov, V.Yu. Makeev and A.B. Rakhmaninova for useful discussions. This study was partially supported by grants from HHMI (55000309) and LICR (CRDF RB0-1268).

References

von Heijne G. Principles of membrane protein assembly and structure // Prog. Biophys. Mol. Biol. 1996. V. 66. P. 113–139.

von Heijne G. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology // EMBO J. 1986. V. 5. P. 3021–3027.

Saier M.H. Jr. A functional-phylogenetic system for the classification of transport proteins // J. Cell Biochem. Suppl. 1999. V. 32-33. P. 84–94.

Sutormin R.A., Rakhmaninova A.B., Gelfand M.S. BATMAS30: amino acid substitution matrix for alignment of bacterial transporters // Proteins. 2003. V. 51. P. 85–95.

Thompson J.D., Higgins D.G., Gibson T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. V. 22. P. 4673–4680.

MOLECULAR DYNAMICS SIMULATIONS FOR LARGE SERIES OF PEPTIDES (COMPARATIVE STUDY)

Shaitan K.V.

M.V. Lomonosov Moscow State University, Moscow, Russia 119992, e-mail: shaitan@moldyn.ru

Keywords: *protein and peptide dynamics, molecular simulations, free energy contour maps, kinematics of conformation transitions*

Summary

Dynamics and physical properties of molecules strongly depend on their potential energy landscapes. Proteins and peptides consist of several definite types natural aminoacid residues and possess some specific dynamics which could be showed during folding. The following questions come about in this context: firstly – if the common features of potential energy surface exist for peptides; secondly – how the protein dynamics corresponds to the individual aminoacid residues dynamics and to the dynamics of the blocks containing a few aminoacids; thirdly – weather the natural aminoacid residues have the original dynamic properties in contrast to similar molecules; and finally – how the surrounding solvents influence on the residues dynamics.

These questions are discussed with the help of methods based on molecular dynamics simulations. The topology of the energy level surfaces for the molecules with conformation flexibility was developed as well.

Results and Discussion

The topology of energy level surfaces for the molecules with conformational mobility was obtained with the aid of Morse theory [1]. The physical idea is similar to the inherent structure approach [2] to glass forming and supercooled liquids. But we use the multidimensional torus as the domain for the definition of the potential energy surface. In this way the representation of energy surface topology is the simplest [3-5]. We should admit that the metric of this conformation space is not Euclidian.

In terms of Morse theory, the hypersurface topology of level E of potential energy U (q) is determined by the behavior of U(q) in the vicinity of critical points (i.e. points where all $\frac{\partial U}{\partial q_i} = 0$). For systems with conformational degrees of freedom the matrix of second derivatives of U (Hessian) at the critical point after reduction to the diagonal form will contain not only positive but also negative elements. The number of negative diagonal elements of diagonalized Hessian at the critical point is called the index of critical point. According to the lemma of Morse in the vicinity of critical point q_0 with index k for N-dimensional surface U (q) (reasonable from physical point of view) there are regular local coordinates $x_i = x_i (q-q_0)$ and eq.(1) is valid

$$U(q)= U(q_0) – x_1^2 - \ldots - x_k^2 + x^2_{k+1} + \ldots + x_N^2 . \qquad (1)$$

In two-dimensional case a simple saddle point is index 1 critical point. In critical points with index 0 the local minima are situated. From eq.(1), it becomes clear that the topology of energy surface E=U(q) is determined by the indexes of critical points and by the relative values of U.

Fig.1 shows the changes in topology of the energy level surface for the rotational isomeric model of an atomic chain. The contribution of torsion angles described by the sum of trigonometric functions $\sin(n_i q_i)$ is of principle here. Thus an ideal lattice assigned to a multidimensional torus arises for the surface of energy level in the space of conformations. Rather small Van der Waals

interactions in hydrocarbons do not badly distort this lattice (except the cases of Van der Waals radii overlapping). Strong Coulomb forces result in lowering of surface symmetry [6]. Some deformations of the energy level surface are also observed for peptides. However the main features of the energy level surface topology are conserved. Fig. 1 illustrates samples the topology of the potential energy (eq. 2) levels surface.

$$U = \sin(n_1 q_1)+\sin(n_2 q_2)+ \sin(n_3 q_3). \quad (2)$$

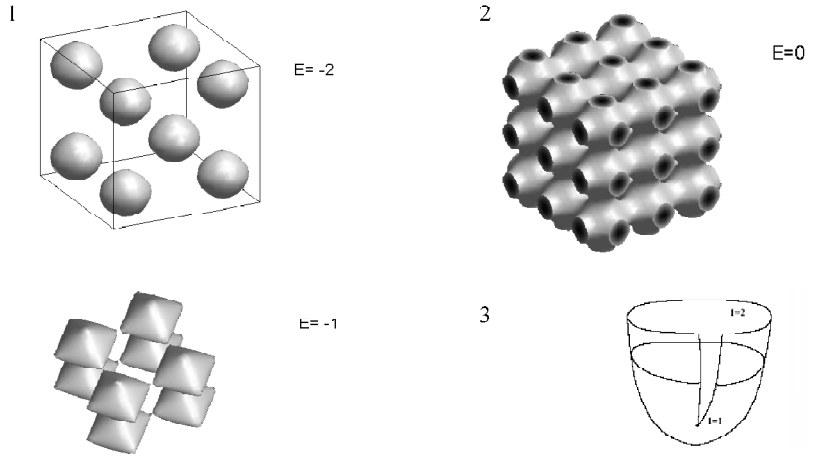


Fig. 1. The energy level surfaces for three conformational degrees of freedom (see eq. (2)) at different values of energy E (in units of the half barrier height). Periodic boundary condition should be applied. (1) – n_i =2, the formation of a simple saddle point at E=-1 (the energy level is enough for overcoming only one barrier); (2) n_i=3, the path or tube through the simple saddle point is broadened, but the energy is still not enough for overcoming two barriers simultaneously. Darker parts denote greater values of kinetic energy. At rising of energy level the dimensions of the paths connecting local minima increase from 0 up to 3 at E=3 (it is not depicted here). At E=0 three-dimensional surface looks like Fermi surface formally. 3 – the scheme of Morse reorganization on the energy level surface. I=1 – a simple saddle point with critical index 1. I=2 – critical point with index 2 (resembles a chemical beaker with deep nose) becomes accessible if energy is enough for overcoming two barriers height simultaneously. Critical points with higher index are hardly presentable in 3D space.

For systems with conformational mobility this topology can be represented as a multitude of hyperspheres (loci) connected by the network of tubes (handles) of different dimensions that pass through multidimensional saddle points. The number and dimension of these tubes determine the scanning rate of energy level surface by representative point.

Kinematics of conformation transitions is defined as rotational displacement of the vectors – $exp[i\phi_k(t)]$, where ϕ_k – torsion angles.

Energy landscapes topography for large series (~10^3) of small peptides in virtual (collision) and water (TIP3P soft model) media is studied by the methods based on molecular dynamics simulations (with AMBER99 force field). We consider the trajectories with good statistics only (T=2000K, trajectory length 10ns). Routines leading to attractors aren't used [7, 8]. Auto- (3) and crosscorrelation (4) $exp[i\phi_k(t)]$ functions, 2-D and 3-D contour maps of free energy (and the results of their cluster analysis) are applied for a classification of the peptides energy landscape topography [9].

$$F(\tau) = \left\langle e^{i\alpha(t)} \quad e^{-i\alpha(t+\tau)} \right\rangle - \left| \left\langle e^{i\alpha(t)} \right\rangle \right|^2 \tag{3}$$

$$F(\tau) = \left\langle e^{i[\alpha(t)-\alpha(t+\tau)]} \quad e^{-i[\beta(t)-\beta(t+\tau)]} \right\rangle - \left\langle e^{i[\alpha(t)-\alpha(t+\tau)]} \right\rangle \left\langle e^{-i[\beta(t)-\beta(t+\tau)]} \right\rangle. \tag{4}$$

The kinematics similarity is observed in a large number of conformational degrees of freedom (see e.g. Fig. 2).

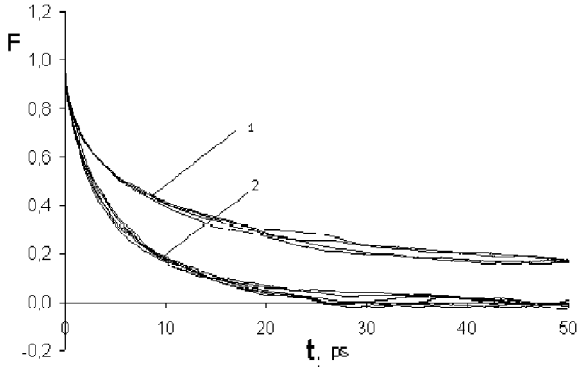


Fig. 2. The real part of the autocorrelation functions (eq.3) of torsion angles in dipeptides. The autocorrelation functions are chosen for: (1) – the angles ψ_1 and φ_2 in dipeptide arg_1-his_2, χ_{21} in glu-phe, χ_{11} in his_1-asn_2 and (2) – the angles ψ_2 in met_1-met_2, φ_1 in gln_1-gln_2, χ_{11} in ser_1-ser_2, ψ_1 in tyr_1-cys_2, φ_1 in cys_1-asn_2, ψ_2 in asp_1-his_2. The dynamic isomorphism for the torsion angles is observed.

An example of autocorrelation functions cluster analysis is presented below for all natural dipeptides (Fig. 3). We should emphasize that in water medium the similarity of torsion angles kinematics in peptides becomes greater.

2-D and 3-D contour maps of free energy and possible symmetry properties of the multidimensional energy surface for natural peptides and chimeras are considered in connection with the funnel topography [10, 11]. The funnel topography and initial conditions effects on folding dynamics of some homopolymer model chains are also simulated.

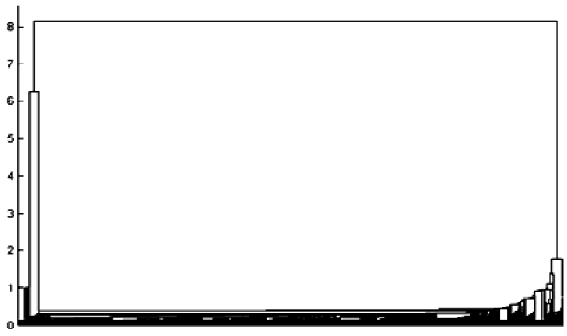


Fig. 3. Cluster analysis data for the kinematics of 2400 torsion angles (abscissa axis) in 400 natural dipeptides (ACE-R1-R2-NHMe) in virtual medium. For the several groups (but not the types) of torsion angles a lot of the autocorrelation functions $exp[if_k(t)]$ are very close (see Fig. 2).

Acknowledgements

This work is supported by RFBR (grant 04-04-49645), Ministry of Education RF and Moscow grant.

References

1. Milnor J.W. Morse theory. Princeton Univ. Press, 1963. 160 p.
2. Debenedetti P.G., Stillinger F.H. Supercooled liquids and the glass transition // Nature. 2001. V. 410. P. 259–267.
3. Shaitan K.V. The topological structure of hypersurfaces of conformational energy levels and physical mechanisms of internal proteins mobility // Macromolecular Symp. 1996. V. 106. P. 321–335.
4. Shaitan K.V. Dynamics of electron-conformational transitions and new approaches to the physical mechanisms of functioning of biomacromolecules // Biophysics (Transl. from Biofizika). 1994. V. 39. P. 993–1011.
5. Shaitan K.V. Energy surface and conformational dynamics of molecules // Russ. J. Electrochem. 2003. V. 39. P. 220–227.
6. Shaitan K.V., Belyakov A.A., Leontiev K.M., Saraikin S.S., Mihailuk M.G., Egorova K.B., Orlov M.V. Energy landscape geometry and conformational dynamics: from hydrocarbons up to peptides and proteins // Chem. Phys. 2003. V. 22. P. 57–68. (In Russ.).
7. Golo V.L., Salnikov Vl.N., Shaitan K.V. Harmonic oscillators in the Nose-Hoover environment // arXiv:cond-mat/0401608 v1 29 Jan 2004.
8. Golo V.L., Shaitan K.V. Dynamic attractor associated with the berendsen thermostat and slow dynamics of biological macromolecules // Biophysics (Transl. from Biofizika). 2002. V. 47. P. 611–617.
9. Shaitan K.V., Ermolaeva M.D., Saraikin S.S. Nonlinear dynamics of the molecular systems and the correlations of internal motions in the oligopeptides // Ferroelectrics. 1999. V. 220. P. 205–220.
10. Shoemaker B.A., Portman J.J., Wolynes P.G. Speeding molecular recognition by using the folding funnel: The fly-casting mechanism // PNAS. 2000. V. 97. P. 8868–8873.
11. Levy Y., Jortner J., Becker O.M. Dynamics of hierarchical folding on energy landscapes of hexapeptides // J. Chem. Phys. 2001. V. 115. P. 10533–10547.

STRUCTURAL AND FUNCTIONAL ANALYSIS OF POORLY CHARACTERIZED PROTEIN FAMILIES AT THE ONTARIO CENTRE FOR STRUCTURAL PROTEOMICS

Skarina T.*, Evdokimova E.*, Yakunin A., Khachatryan A., Pennycooke M., Guido V., Guthrie J., Xu X., Semesi A., Gu J., Kudritska M., Egorova O., Gorodichtchenskaia E., Yee A., Savchenko A., Arrowsmith C.H., Edwards A.M.

Ontario Center for Structural Proteomics, University of Toronto, University Health Network, 112 College Street, Toronto, ON, M5G 1L6, Canada

***** Corresponding authors: e-mail:tskarina@uhnres.utoronto.ca, evdokimo@ uhnres.utoronto.ca

Keywords: *structural proteomics; three-dimentional protein structure; protein cloning, purification, crystallization*

Summary

Motivation: Consensus strategy for structural proteomics is to determine the experimental structures for enough proteins such that remaining structures can be predicted accurately using computational approaches. This approach should eventually lead to completion of the protein folding space and elucidation on the possible functions for these proteins. Expected benefits also include faster identification of new structure-based medicines, improved therapeutics for diseases treating and development of technology and methodology for protein production and crystallography.

Results: In last three years our pipeline produced more that 60 structures of novel protein families representatives. This information was used for accurate structure prediction/modeling of several hundred homologues of these proteins.

Availability: Information on our results can be found on the web site: https://www.uhnres.utoronto.ca/proteomics/; https://www.uhnres.utoronto.ca/proteomics/db/index_summer.php (access available on request)

Introduction

The Ontario Center for Structural Proteomics is a Canadian and NIH-funded Center affiliated with both the Northeast Structural Genomics Consortium and the Midwest Center for Structural Genomics in the US. These consortiums are working to establish the pipeline for production of hundreds novel protein structures per year.

Methods

Our main methods to determine three-dimentional protein structures are X-ray crystallography and NMR spectroscopy. Our methods for structural biology sample preparation include target selection (Bertone *et al*., 2001), cloning, protein expression, and solubility evaluation, protein purification and crystallization (Savchenko *et al*., 2003).

Implementation and Results

Our group selects protein targets from all three kingdoms of life (Bacteria, Archaea and Eukarya), with an emphasis on previously unknown folds and on proteins from disease-causing organisms that are more challenging for structural studies. One of the main criteria for selection is the absence of strong (less that 30 % ID) sequence similarity to proteins with known 3D structure deposited into PDB. Our project has already covered more than 3000 structurally uncharacterized proteins from *Escherichia coli*, *Thermotoga maritima*, *Pseudomonas aeruginosa*, *Thermoplasma*

acidophilum, *Methanobacterium thermoautothrophicum*, *Saccharomyces cerevisiae* and other completed genomes.

We developed and standardized the procedure for high throughput protein purification by affinity chromatography, which permitted us to purify 10–12 different proteins in parallel. We have also developed a pipeline for screening for crystallization conditions (Kimber *et al*., 2003), which allowed us to obtain initial crystallization conditions for 30–40 % of purified proteins. Although not completely automated this strategy remains one of most efficient in the field of structural proteomics. We also developed high-throughput enzymatic assays to identify novel enzyme among uncharacterized proteins that we purify for structural studies (Yee *et al*., 2002).

Discussion

We continue to develop the technology and methodology for large-scale, high through

put structural biology sample preparation and the protein structure solving.

Our technological aims are:

- effective valid protein target identification in sequenced genomes;
- development of highly parallel and cost-effective methods to clone, express and purify proteins on a scale required for structural studies;
- improvement of protein crystallization process by automatization and creating more effective crystallization screens;
- acceleration of 3D protein structures solving process;
- development of effective database;
- testing and using of all appropriate novelty developments in computational analysis and robotization.

Acknowledgements

Our group benefits from collaborations with all investigators of the NESG and MCSG as well as many biology labs in Canada and around the world. In particular, the following laboratories have solved 3D structures for crystals or NMR samples produced in our Center: Joachimiak A., Korolev S., Kim Y., Zhang R., Sanishvili R. (ANL); Kennedy M. (PNNL); Pai E. (U. of Toronto); Tong L., Hunt J. (Columbia U.); Montelione G. (Rutgers U.); Gehring K. (McGill U.); McIntosh L. (U. of British Columbia); Lee W-T. (Yonsei U.); Wishart (U. of Alberta).

References

Bertone P., Kluger Y., Lan N., Zheng D., Christendat D., Yee A., Edwards A., Arrowsmith Ch., Montelione G., Gerstein M. Spine: an integrated tracking database and data mining approach for identifying feasible targets in high-throughput structural proteomics // Nucl. Ac. Res. 2001. V. 29, N 12. P. 2884–2898. Oxford University Press.

Kimber M., Vallee F., Houston S., Necakov A., Skarina T., Evdokimova E., Beasly S., Christendat D.,Savchenko A., Arrowsmith Ch., Vedadi M., Gerstain M., Edwards A. Data Mining Crystallization Databases: Knowledge-Based Approaches to Optimize Protein Crystal Screens // Prot. 2003. V. 51. P. 562–568.

Savchenko A., Yee A., Khachatryan A., Skarina T., Evdokimova E., Pavlova M., Semesi A., Northey J., Beasley S., Lang N., Das R., Gerstein M., Arrowmith Ch., Edwards A. Strategies for structural proteomics of prokaryotes: Quantifying the advantages of studying orthologous proteins and of using both NMR and X-ray crystallography approaches // Prot. 2003. 15; 50(3), 392-9 PMID: 12557182.

Yee A., Pardee K., Christendat D., Savchenko A., Edwards A., Arrowsmith Ah. Structural Proteomics: Toward High-Throughput Structural Biology as a Tool in Functional Genomics // Acc. Chem. Res. 2003. V. 36. P. 183–189.

A MARKOV MODEL FOR PROTEIN SEQUENCES

Surya pavan Y., Mitra Chanchal K.*

Department of Biochemistry, University of Hyderabad, Hyderabad –500 046
* Corresponding author: e-mail: ckmsl@uohyd.ernet.in; surya_pavan@yahoo.com

Keywords: *Markov process, Anova test, hidden Markov model*

Summary

The aim was to see whether amino acids in the SWISS-PROT protein sequence database follow any specific pattern.

Motivation: Understanding the behavior pattern of amino acids in protein sequences may be useful in predicting protein structure as well as its function. However, before we analyze the behavior of amino acids one has to show that the amino acids follow a steady and stationary distribution in the protein sequences. This can be tested effectively by using Markovian models.

Preliminary studies on protein sequences suggest that there is no clear order or correlation between amino acids in protein sequences. However, careful investigations show that there exists a clear correlation between the amino acids residues in a database. However, most databases cannot be considered random or independent and conclusions derived from studies based on a biased database can be fallacious.

Our earlier works in this direction show that (i) there is a clear correlation between the amino acids in a protein sequence and (ii) the distribution of amino acids in the database follows a fractal pattern. Based on these observations, we can classify protein sequences as multifractals [1].

In this presentation, we report on our investigation in the Markov model as applied to protein sequences. We consider that protein sequences may be considered as a potential candidate for a hidden Markov process.

Introduction

Functionally, proteins are the most diverse of all biological macromolecules. Proteins provide structural support, catalyze cellular reactions, act as signaling molecules, act as channels and transporter and carry out a myriad of other tasks. All proteins, whether from the most ancient lines of bacteria or from the most complex forms of life, are constructed from the same ubiquitous set of 20 amino acid [2]. The amino acids almost never occur in equal amounts in proteins. Although the basic components of all proteins are the same 20 amino acids, they differ in the order or sequence from one protein to another. Understanding the order of the amino acids in a protein sequence and its relation to structure and function can be a baffling problem. In our work, we have tried to show that there are some rules to follow by which one can achieve the problem. We assumed that the protein sequences follow a Markov process and tried to see whether such an assumption can be supported. In this paper, we show evidences to support this view.

For our computations, we have used the Swiss-Prot protein sequence databank Release 37, 1999. We have ignored all fragments (partial sequences) and short sequences (less 256). The final database (working database) has 41,408 sequences with 22,408,660 residues. All the computations have been carried out on an IBM compatible PC using C++ under Linux (gcc).

Model

To show that the amino acids follow a stationary distribution we used two methods.

I. Distribution of amino acids along the sequence attains a stationary value, consistent with a Markov process.

II. The transition matrix shows the expected behavior.

I. Distribution of amino acids along the sequence length

We have computed the distribution of the 20 amino acid residues along the sequence length. The Figure shows the results obtained for Alanine and Cysteine (upto 100 position). It can be seen clearly that the distribution is not uniform in the initial regions of the sequences but stabilizes thereafter.

To verify whether these differences are statistically significant we have carried out an analysis of variance [3] (ANOVA) on these results as detailed below.

Anova. We have divided the protein database into 8 equal databases for comparison. Each mini database consists of 5176 protein sequences. We have selected regions of 10 to 20, 50 to 60, 100 to 110, 150 to 160 and 200 to 210 positions in the protein sequences and counted the frequency of the 20 amino acids at the selected positions (i.e., 10–20, 50–60, 100–110, 150–160 and 200–210). The eight frequencies for each of the 20 amino acids, corresponding to eight different databases, each at 5 different positions, are now compared.

The obtained frequencies from all the eight databases were normalized (converted to per million) and used for comparing the distributions among the selected positions. The frequency counts were converted to per million by using the formulae:

$$\frac{\text{Actual count of amino acid residue at selected region} * 1.0\text{E} + 6}{\text{Total counts of all amino acids in the mini database upto 256th position}}$$

We now compare the normalized frequencies to test for differences.

The bias present in the Swiss-Prot protein sequence databank can be attributed to several factors, but it is certain that the database cannot be considered a random sample. The ASTRAL SCOP (http://astral.berkeley.edu) suggests a novel idea in selecting a set of non-redundant, and possibly a random representative, protein sequence database. We have chosen a set of protein sequences with less than 40 % and 95% sequence similarity for our reference. Therefore, the same computations have been repeated with this non-redundant database also.

II. Comparison of the two distributions

To show any discrimination between the distributions of amino acids at the selected regions we have used log-odds ratio method [4].

$$S(x) = \log \frac{P(x \mid \text{model} +)}{P(x \mid \text{model} -)} = \sum_{i=1}^{L} \log \frac{a^{+}_{x_{i-1}x_i}}{a^{-}_{x_{i-1}x_i}} = \sum_{i=1}^{L} \beta_{x_{i-1}x_i} ,$$

where x is the sequence and $\beta_{x_{i-1}x_i}$ are the log likelihood ratios of corresponding transition probabilities.

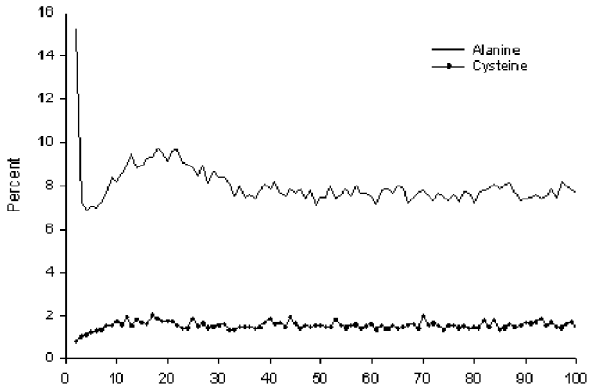
Model + = proposed non-stationary Markov chain (beginning regions of the protein sequences in SWISS-PROT).

Model - = proposed stationary Markov chain (any middle or end regions of protein sequences in SWISS-PROT).

Results and Discussions

We have earlier made attempts to show that the amino acids distribution in protein sequences follow a Markov process [5]. However, we showed it by using very effective statistical methods. The database used for analysis has 77,976 sequences with 22,408,660 residues. We have ignored fragments and short sequences (less than 256 residues). We have used this to avoid one-sided bias

in the results. Initially, we have studied the positional distribution of amino acids in the SWISS-PROT database and the Figure is shown below.



From Table 1 it is evident that the F values obtained between 10–20 with 50–60 have shown more than the critical value proving that the regions compared show difference in distribution. The obtained F value of Alanine in Table 1 shows that there is difference in the distribution of 10–20 with 50–60 regions, which is a clear evident from the graph above. Similarly, the Cysteine distribution showed a steady state in the graph and the F values of the Cysteine prove that they follow stationary distribution altogether in the database. However, in Table 1 the F values obtained

Table 1. The F values of all the common amino acids after comparing the selected regions by using ANOVA method

	10-20 with 50-60	50-60 with 100-110	100-110 with 150-160	150-160 with 200-210
A	18.2706164	0.0220103	0.3061278	0.9853327
C	1.9183966	1.3487991	0.0063689287	0.0145383
D	41.4538523	1.4586969	0.1882551	0.2389655
E	9.3971915	0.0274597	0.4172285	0.3123983
F	0.1608799	1.0700434	0.2132119	0.5935168
G	0.0656247	0.0652373	0.1116950	0.0507178
H	16.0717254	0.6826251	0.4396801	1.3337909
I	0.6754037	0.0219082	2.3722271	0.0230542
K	6.0200729	0.3865508	0.1495042	0.3280580
L	54.5066549	0.5236030	0.0004631	0.001790002
M	5.6762705	4.3139078	0.1553002	0.0131109
N	5.6991288	1.5261917	1.8461342	0.7399015
P	1.8058413	0.0238219	0.0491274	0.0532888
Q	1.0115292	3.6845411	0.0158412	0.00492970
R	0.3638405	0.9887972	0.3182467	0.0818374
S	15.9431021	2.9277288	0.8740657	0.9825878
T	1.5371838	4.1781449	0.4429651	0.3004622
V	0.4225060	1.2997469	0.0259056	0.3799273
W	0.8945887	2.9110155	0.9476149	2.9993862
Y	18.2153837	3.9017292	0.3928937	0.00511537

after comparing the middle and end regions show below the critical value proving that they follow a stationary distribution. The SWISS-PROT database that we considered for our analysis is a redundant database. The difference in the distribution at the beginning regions comparing with the middle and end regions could be because of the redundancy in the database. We then analyzed the ASTRAL SCOPE non-redundant database for distribution patterns and found that the amino acids follow a stationary distribution in all the selected regions both at the beginning and in end regions.

To further substantiate that the selected 50 to 60 regions in SWISS-PROT follow a stationary distribution we have checked the following formulae [6]:

$$\mu_i \mathrm{P}_{\mathrm{ij}} = \mu_j$$

where μ_i is the probability value of the amino acids at the selected region (50–60). P_{ij} is the transition matrix of the selected region. As the product of μ_i and P_{ij} is nearly equal to μ_i proving that the distribution follows a stationary distribution (Table 2).

Table 2. The table shows the probability values at 50 to 60 region (μ_i), which is approximately similar with the μ_j values obtained, proving that the distribution follows a stationary distribution

Amino acids	μ_i	μ_j
A	0.0767219	0.0761882
C	0.0154723	0.0147242
D	0.0535918	0.0551777
E	0.0638650	0.0628719
F	0.0396964	0.0409486
G	0.0717011	0.0716263
H	0.0239028	0.0231501
I	0.0566297	0.0583100
K	0.0585835	0.0576869
L	0.0928366	0.0929337
M	0.0197830	0.0198343
N	0.0438259	0.0437935
P	0.0502037	0.0488722
Q	0.0411213	0.0405356
R	0.0513702	0.0515818
S	0.0728435	0.0722034
T	0.0578324	0.0585128
V	0.0671998	0.0675232
W	0.0120407	0.0115388
Y	0.0306815	0.0319214

Acknowledgements

The work reported above has been made possible by a grant from the University Grants Commission (UGC) and from the Department of Science and Technology (DST) of the Government of India.

References

1. Barnsley M.F. Fractals Everywhere. Academic Press, 1988. P. 172–206.
2. Lehninger A.L., Nelson D.L. Principles of Biochemistry. Second edition, Worth Publishers, New York, 1993.
3. Sokal R., Rohlf F.J. Introduction to Biostatistics. P. 164.

4. Durbin R., Eddy S., Krogh A., Mitchison G. Biological Sequences analysis. Cambridge University press, 1998. P. 51.
5. Rani M., Mitra C.K. Pair-Preferences: a quantitative measure of regularities in protein sequences // J. Biomolecular Structure and Dynamics. 1996. V. 13. P. 935–944.
6. Cover Th.M., Thomas J.A. Elements of Information Theory. 1991. P. 67.

MOLECULAR MODELING OF HUMAN MT_1 AND MT_2 MELATONIN RECEPTORS

Tchugunov A.O.[1,2], *Chavatte P.*[3], *Efremov R.G.*[1]

[1] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, ul. Miklukho-Maklaya 16/10, GSP Moscow, 117997 Russia; [2] Department of Bioengineering, Biological Faculty, M.V. Lomonosov Moscow State University, Vorobiovy gory, 119899, Moscow
[3] Faculte des Sciences Pharmaceutiques et Biologiques - BP 83 - 59006 Lille Cedex - France
* Corresponding author: e-mail: volster@nmr.ru

Keywords: *GPCR, homology modeling, molecular docking, hydrophobic complementarity*

Summary

Motivation: A great number of diseases are related to malfunction of the G-protein coupled receptors (GPCR). Molecular modeling of these receptors is believed to be extremely useful in pharmacology – for design of new drugs with high affinity and specificity.

Results: We propose three-dimensional models of human MT_1 and MT_2 melatonin receptors. The models are further employed to study binding of melatonin and a variety of its analogs to the receptors' active sites. An efficient computational procedure destined to optimization of the ligand binding to the receptors' active site is developed. It includes rotation of one of transmembrane (TM) α-helices around its axis followed by simultaneous assessment of the quality of the resulting complexes according to a number of criteria. The last ones were elaborated in the result of statistical analysis of a large set of high-resolution X ray structures of protein-ligand complexes.

Introduction

GPCRs represent a very important class of integral membrane proteins. They share a common folding motif – seven TM α-helices. The binding site for small organic molecules such as monoamines, nucleotides, short peptides etc., is located in the membrane protein domain, in a cavity formed by TM α-helices.

Melatonin, a derivative of tryptophan, is an important biological agent, responsible for circadian rhythm regulation. It also possesses immunomodulator and antioxidant activities. Building of molecular models of MT_1 and MT_2 receptors is necessary for design of high-specific and high-affine agonists and antagonists of melatonin. These compounds are believed to be very useful as prototypes for new perspective drugs.

Methods

Molecular models of MT_1 and MT_2 receptors were built using homology between their amino acid sequences and that of bovine rhodopsin, whose 3D structure has been determined by X ray crystallography [1]. This was done using the MODELLER [2] software. The hydrophobic organization of the models was assessed using the molecular hydrophobicity potential approach. The variability properties of TM helices were calculated based on sequence alignment for MT_1, MT_2 and homologous receptors. Both starting models and the conformers with rotated TM3 segment were optimized using a multi-step energy minimization protocol. At this stage, a number of experimentally derived distance restraints between protein and ligand atoms were employed. Resulting 3D models were used for molecular docking with melatonin and a set of its analogs. This was done with the GOLD [3] software. Finally, the optimal complexes were selected using a novel criterion that is based on complementarity of hydrophobic properties between the ligand and the receptor.

Results

Analysis of hydrophobic and variable properties of TM helices in the melatonin receptors leads to the following conclusions. The most hydrophobic side of each helical segment faces the lipid environment, while the most hydrophilic one is buried inside the protein. The vectors of variability moments calculated for each helix, point to the exterior of the bundle. These results are very similar to those inherent in rhodopsin. A proposal was made that the hydrophobic and the variability properties of the membrane domains are well conserved in the GPCR family.

From mutagenesis studies it is known that melatonin binds to MT_1 and MT_2 receptors via hydrogen bonds to Ser110/123, Ser114/127 in TM3 and His195/208 in TM5, respectively. This data were used as restraints in the molecular docking protocol. However, it should be noted that the structural template provided by the dark-adapted rhodopsin may be not well-suited to accommodate melatonin in the active site. To explore conformational possibilities of the ligand in its bound state, the following computational procedure was proposed. TM3 helix of the model was rotated in a clockwise direction (as seen from the extracellular side) around its axis. In total, 12 variants of the model for each receptor were generated. Quality of the models was estimated based on analysis of the ligand-receptor hydrogen bonding and the GOLDSCORE function. As a result, two optimal conformations of the complexes were chosen for future analysis and consideration. According to the experimental data, both of them accommodate melatonin and its analogs in the best way.

These "optimal" models of MT_1 and MT_2 were used in exhaustive molecular docking simulations with melatonin and its analogs. Analysis of docking results was carried out using a novel criterion to select the most probable ligand orientations in the binding site. Thus, complementarity between hydrophobic properties of the ligand and the receptor was taken into account in a quantitative manner. Such a criterion was initially developed and tested via statistical analysis of a large number of high-resolution X ray structures of different protein-ligand complexes.

The results are currently being used to rationalize available experimental data and to design new experiments on site-directed mutagenesis of the receptors. Also, the data obtained will be employed to goal-oriented design of new analogs of melatonin with high affinity and selectivity.

Acknowledgements

This work was supported in part by the Programme RAS MCB and the Ministry of Science and Technology of Russian Federation (the State contract No. 43.073.1.1.1508), and by the Russian Foundation for Basic Research (grant 04-04-48875a).

References

1. Marti-Renom M.A., Stuart A., Fiser A., Sánchez R., Melo F., Šali A. Comparative protein structure modeling of genes and genomes // Annu. Rev. Biophys. Biomol. Struct. 2000. V. 29. P. 291–325.
2. Nissink J.W., Murray C., Hartshorn M., Verdonk M.L., Cole J.C., Taylor R. A new test set for validating predictions of protein-ligand interaction // Proteins. 2002. V. 49. P. 457–471.
3. Palczewski K., Kumasaka T., Hori T., Behnke C.A., Motoshima H., Fox B.A., Le Trong I., Teller D.C., Okada T., Stenkamp R.E., Yamamoto M., Miyano M. Crystal structure of rhodopsin: a G protein-coupled receptor // Science. 2000. V. 289. P. 739–745.

LATENT PERIODICITY OF THE PROTEIN FAMILIES

Turutina V.P.*,Korotkov E.V., Laskin A.A.

Bioengineering Center of RAS, Moscow, Russia, e-mail: veratp@yandex.ru

Keywords: *latent periodicity, alignment, information decomposition, noise decomposition, profile analysis, repeats*

Summary

Motivation: The sequences of proteins possessing strongly divergent periodicity are especially poorly studied at present. Using the developed original methods we have investigated the presence of the latent periodicity in amino acid sequences of various proteins and its connection with evolutionary formation of genes and protein functions.

Results: We have revealed 100 protein families for which from 70 up to 100 % of the proteins included in them possess the same kind of the latent periodicity. The periodicity of the certain length and type is a characteristic of each family. We suppose that latent periodicity of functional domains in proteins has the common character and can be found for all known protein families.

Availability: available on request from the authors.

Introduction

Periodicity is revealed only for rather small number of proteins today (Katti *et al*., 2000). All repeats found in amino acid sequences can be divided in three groups (Marcotte *et al*., 1999): 1) consecutively duplicated independent functional and structural units which can function independently from each other (for example, the domains arranged by a principle of zinc "fingers"); 2) well appreciable repetitions which make up the common functional subunit, but are not functional separately (examples are well-known "b-barrel" pattern, leucine-rich repetitions and tetratricopeptide repeats); 3) the least investigated group of the amino acid repetitions which have no significant internal homology. This group of repeats can be named repeating motives as it is usually possible to detect a regularity only in position of single amino acid residues, as in a case of leucine residue in the leucine zipper (Landschulz *et al*., 1988), or a regularity in distribution among the period positions oof the amino acid residues possessing common properties, e.g. hydrophilic or hydrophobic.

Earlier the method of searching for the latent periodicity (Laskin *et al*., 2003) has allowed to find out the feebly marked or latent periodicity in more than 80 % NAD-binding sites, in active sites Serine-threonine and Tyrosine protein kinases, and also in 16 various protein domains (Laskin *et al*., 2004a, b).

Based on the results received, in the current paper we have tried to find out how common is the phenomenon of the presence of latent (or feebly marked) amino acid periodicity in the active centers of various proteins and in protein domains. For this purpose we have completely analyzed Swiss-prot databank (release 41) using the method of searching for the latent periodicity having the higher sensitivity.

Methods

Methods of searching for the latent periodicity in amino acid sequences of proteins used by us (Korotkov *et al*., 2003; Laskin *et al*., 2003) allow to see so strongly diverged repeats in proteins, that they are not visible to the naked eye and are determined exclusively due to statistical features in distribution of amino acids in protein sequence.

For the purpose of revealing of the latent periodicity in protein domains we have analyzed all

Swiss-prot databank (release 41) using the method of Information Decomposition (ID) (Korotkov *et al*., 2003) to search for the latent periodicity in amino acid sequences without taking into account insertions and deletions of symbols. Based on found (initiating) amino acid sequences, matrices of periodicity were defined. Elements of such matrices showed the occurrence of each amino acid in every position of the period for all sequence. Then each matrix of this kind was used to reveal the latent periodicity in view of possible insertions and deletions of symbols. To perform the further profile analysis, the elements of the corresponding position-weight matrix *W* were calculated using the following formula:

$$W_{i,j} = C \ln \frac{p_{i,j}}{f_i}, \tag{1}$$

where $W_{i,j}$ – an element of a position-weight matrix for symbol *i* in a position *j*, $p_{i,j}$ – occurrence frequency of symbols such as *i* in a position *j*, and f_i – occurrence frequency of symbols such as *i* in amino acid sequences where the latent periodicity has been revealed.

We then used a method of Noise Decomposition (ND) (Laskin *et al*., 2003) to generate the position-weight matrix of periodicity *W*, which could be used to reveal the latent periodicity in all protein domains of one family at statistically significant level and to not reveal similarity in all other amino acid sequences. The essence of the iterative ND method is:

1. At the first stage we search for amino acid sequences from Swiss-prot databank (release 41) which have statistically significant periodicity defined by the position-weight matrix W according to the formula (1). In this search we used the modified profile analysis (Laskin *et al*., 2004a; Chaley *et al*., 2003). As a result of scanning sequences, the sequences that are aligned optimally concerning the matrix *W* have been selected.

2. At the second stage we divided the selected sequences into two sets. The first one included sequences with the same functional value as initiating amino acid sequence. This set can be referenced as "true positive". All others were included in the second set of amino acid sequences. We have named this set "false positive".

3. At the third stage we modified position-weight matrix *W* for two purposes. First, to find out as many as possible similarities related to the "true positive" set having the significance value "Score" greater than 6.0 (statistically significant level of Score). The second purpose was to reduce and, whenever possible, to reduce to zero the number of similarities with the significance value Score greater than 6.0 in the "false positive" set. The matrix $W_{i,j}$ was modified by the formula:

$$\overline{W}_{i,j} = C \ln \frac{r_{i,j}}{\pi_{i,j}} \tag{2}$$

For calculation of $r_{i,j}$ we made the global pairwise alignment for all sequences of the "true positive" set and calculated the weight of global alignment as *S (k, l)* in each case. Then we have introduced the *T (k)* value, which shows the representation of the given sequence in the "true positive" set.

$$T(k) = \sum_l \max(0, S(k,l) / \{\max(S(k,k), S(l,l))\}). \tag{3}$$

Then we calculated $r_{i,j}$ as:

$$r_{i,j} = \sum_k p^k{}_{i,j} / T(k), \tag{4}$$

where $p^k{}_{i,j}$ are similar to probabilities $p_{i,j}$ calculated for k^{th} sequence.

Value $\pi_{i,j}$ was defined as:

$$\pi_{i,j} = c_0 f_i + c_1 \sum_k q^k{}_{i,j} / N_1 . \quad (5)$$

Probabilities $q^k{}_{i,j}$ are similar to probabilities $p^k{}_{i,j}$, but they have been determined only for k^{th} sequence from the "false positive" set. N_1 is a potency of the "false positive" set of sequences. Probabilities f_i are defined as frequencies of amino acids in the "false positive" set of amino acid sequences.

Results and Discussion

We have revealed the latent periodicity in 100 protein families. And from 70 up to 100 % of proteins in each family possess the appropriate latent periodicity. As an example, the following values are shown for 10 protein families in Table: period length of the revealed periodicity, full number of proteins belonging to the considered family in Swiss-prot databank, and also number of proteins in family in which the latent periodicity has been found.

Table. 10 families of protein domains from 100, possessing the latent periodicity

	Name of protein family	Period length (aa)	Number of proteins in the protein family	Number of true positive
1	Homeobox protein (homeobox domain)	14	725	572
2	MADS box protein (domain MADS)	13	73	70
3	T-Box protein (T-box domain)	14	65	59
4	P450-protein (chain cytochrome P450, act_site)	14	665	577
5	PyruvateKinase (ADP binding site)	11	67	59
6	Protein Cpn10 (subunit hasn't been marked out)	14	133	121
7	CF(1) – ATP synthase beta chain	7	135	134
8	CF(1) – ATP synthase alpha chain	9	92	89
8	Lysozyme C (chain lisozyme C, active site)	6	64	60
9	Phorbol-ester binding (domain phorbol-ester and dag binding)	36	108	88
10	Chalcone Synthase (subunit hasn't been marked out)	17	119	117

It is worth noting that the number of protein families and simple domains possessing the latent periodicity apparently is much more than 100. Within the framework of the given work we have stopped the search upon finding the100 families though, in principle, searching can be continued further. It seems to us that the size of the latent periodicity data received for the protein families having various functional and biological value, together with earlier received results (Laskin *et al*., 2003; Laskin *et al*., 2004b), shows, that we do deal with the biological phenomenon which is common enough.

The results received support a hypothesis that the origin of primary genes is the plural tandem duplications and show that the periodicity of amino acid sequences can be specific for the biological functions of proteins. It is very probable, that in the present work we have revealed the traces of gene formation processes.

At the same time our data show that each latent period defines the corresponding secondary structure of amino acid sequence. Therefore we can assume that the characteristic periodicity of element secondary structure is intrinsic for the domain with the certain biological function. The results received by us show that the certain accordance exists between the period lengths of amino acid sequences and functional meaning of protein families where the latent periodicity has been found. If such accordance really takes place for all protein families, it can be found using the methods ID and ND developed by us.

References

Chaley M.B., Korotkov E.V., Kudryashov N.A. Latent Periodicity of 21 bases TypicaL for MCP II Gene is Widely Present in Various Bacterial Genes // DNA Sequence. 2003. V. 14. P. 37–52.

Katti M.V., Sami-Subbu R., Ranjekar P.K., Gupta V.S. Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications // Protein Sci. 2000. V. 9. P. 1203–1209.

Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method to analyze symbolical sequences // Phys. Let. A. 2003. V. 312. P. 198–210.

Landschulz W.H., Johnson P.F., McKnight S.L. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins // Science. 1988. V. 240. P. 1759–1764.

Laskin A.A., Korotkov E.V., Chaley M.B., Kudryashov N.A. The locally optimal method of cyclic alignment to reveal latent periodicities in genetic texts: the NAD-binding protein sites // Mol. Biol. 2003. V. 37, N 4. P. 561–570.

Laskin A.A., Kudryashov N.A., Skryabin K.G., Korotkov E.V. Revealing the latent periodicity in protein domains by a iterated profile analysis JMM, 2004a. in press.

Laskin A.A., Korotkov E.V., Kudryashov N.A. Latent periodicity of many domains in rotein sequences reflects their structure, function and evolution // Bioinformatics of genome regulation and structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer press, 2004b. P. 135–144.

Marcotte E.M., Pellegrini M., Yeates T.O., Eisenberg D. A census of protein repeats // J. Mol. Biol. 1999. V. 293. P. 151–160.

Author index

Keywords

Подготовлено к печати
в редакционно-издательском отделе ИЦиГ СО РАН

Подписано к печати 23.06.2004
Формат бумаги 70x108/16. Усл.-печ.л. 43,34. Уч.-изд.л. 38,1
Тираж 250. Заказ 335
Отпечатано в типографии СО РАН
630090, Новосибирск, Морской просп., 2