RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

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

Volume 3

BGRS' 2002
Novosibirsk, Russia
July 14 - 20, 2002

IC&G, Novosibirsk, 2002

International Program Committee

Nikolay Kolchanov, Institute of Cytology and Genetics, Novosibirsk, Russia (*Chairman of the Conference*)
Ralf Hofestadt, University of Bielefeld, Germany (*Co-Chairman of the Conference*)
Philip Bourne, SDSC, San-Diego, USA (*Co-Chairman of the Conference*)
Nickolai Alexandrov, Ceres Inc., Malibu, USA
Philipp Bucher, Swiss Institute for Experimental Cancer Research, Switzerland
Julio Collado-Vides, National University of Mexico, Mexico
Jim Fickett, AstraZeneca, Boston, USA
Paolo Frasconi, University of Florence, Firenze, Italy
Sergey Goncharov, Sobolev Institute of Mathematics, Novosibirsk, Russia
Igor Goryanin, GlaxoSmithKline, UK
Charlie Hodgman, GlaxoSmithKline, UK
Elza Khusnutdinova, Institute of Biochemistry and Genetics, Ufa Sci. Centre RAS (Ufa), Russia
Lev Kisselev, Engelhardt Institute of Molecular Biology, Moscow, Russia
Boris Kovalerchuk, Central Washington University (Ellensburg), USA
Luciano Milanesi, ITBA, Milan, Italy
John Reinitz, The University at Stony Brook, N.Y., USA
Akinori Sarai, RIKEN Tsukuba Life Science Center, Tsukuba, Japan
Ilya Shindyalov, San Diego Supercomputer Center, USA
Rustem Tchuraev, Institute of Biology, Ufa Sci. Centre RAS, Ufa, Russia
Masaru Tomita, Institute for Advanced Biosciences, Keio University, Japan
Edgar Wingender, GBF, Braunschweig, Germany
Nikolay Yankovsky, Institute of General Genetics, Moscow, Russia
Lev Zhivotovsky, Institute of General Genetics, Moscow, Russia

Local Organizing Committee

Dagmara Furman, Institute of Cytology and Genetics, Novosibirsk,
Nadya Omelianchuk, Institute of Cytology and Genetics, Novosibirsk,
Sergey Lavryushev, Institute of Cytology and Genetics, Novosibirsk,
Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk,
Elena Borovskikh , Institute of Cytology and Genetics, Novosibirsk,
Nikolay Shkel, Institute of Cytology and Genetics, Novosibirsk,
Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk,

The information about the Conference BGRS' 2002 is presented at http://www.bionet.nsc.ru/meeting/bgrs2002/

Our sponsors

Organizers

Institute of Cytology and Genetics, SB RAS

Siberian Branch of the Russian Academy of Sciences

Grants

INTAS Conference Grant

Glaxo Wellcome Inc.

Russian Foundation for Basic Research

Ministry of Industry, Science and Technologies of the Russian Federation

Information sponsors

San Diego Supercomputer Center, United States

RIKEN Tsukuba Institute

Bielefeld University, Faculty of Technology

http://www.karger.com/

In Silico Biology

Others

KWESTA-group: computers, computer accessories, service

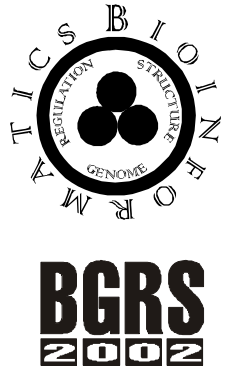
CONTETS

RNA COMPUTATIONAL BIOLOGY

COMPUTATIONAL PROTEOMICS

METHODOLOGICAL PROBLEMS OF BIOINFORMATICS

OTHER TOPICS RELATED TO BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE



RNA COMPUTATIONAL BIOLOGY

GARNA INTERNET RESOURCE FOR THE ANALYSIS OF RNA SECONDARY STRUCTURE: ITS STATE IN 2002

**** Vorobiev D.G., Titov I.I., Ivanisenko V.A.***

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

Key words: *RNA, secondary structure, genetic algorithm, Internet*

Summary

Motivation: Application of genetic algorithms (GAs) for prediction of RNA secondary structure (SS) opens up fresh opportunities because they are capable of taking into account the energy of tertiary interactions and reveal kinetic SS intermediates.

Results: Here we describe the present state of the Internet resource GArna for SS RNA analysis. We use GA for predicting RNA structures online. A separate block allows estimation of the evolutionary significance of the predicted structure on the base of the Z-score statistics. The resource also includes a block for calculating the context index of E-score illustrating the potential of SS formation.

Availability: http://www.mgs.bionet.nsc.ru/mgs/programs/2dstructrna/

Introduction

Calculation of RNA secondary structure is one of the earliest tasks in bioinformatics. The most common approach is minimization of the free energy of the structure. The mfold algorithm by Mathews et al. (1999) allows calculation of several structures of lowermost energies. The algorithm of statistical sums calculates the probabilities of formation of complementary pairs (McCaskill, 1990). The kinetic approach simulates formation of RNA SS (Mironov et al., 1986). Of the listed algorithms, only mfold is available from Internet. In recent years, the genetic algorithm has been used for analysis of RNA structures (Benedetti, Morosetti, 1995; Currey, Shapiro, 1997; Gultyaev et al., 1995; Proutski, 1997; Titov et al., 2002). One of its advantages is that it can find intermediate states of RNA folding. In this feature, it is similar to kinetic algorithms and the gradient descent approach. However, GA has a feature enhancing its optimization ability. This feature is recombinations. Its essence is that a combination of blocks of two "good" solutions may yield even better one. Another important advantage of GA is that it allows taking into account tertiary interactions, e.g., pseudoknots (Gultyaev et al., 1995).

This study describes the present state of the resource GArna we developed for RNA SS analysis (Titov et al., 2002). It includes the pioneering GA for online prediction of RNA SS and blocks for calculating some helpful indices of RNA SS.

Methods and Algorithms

I. GA for predicting RNA SS

Let us outline the protocol of GA operation:

1. Construction of a set {h} of all possible stems for the RNA molecule under study. Incomplete helices, whose combinations give rise to "running loops", are allowed. Energy is calculated for each stem (Jaeger et al., 1989).

2. Construction of the starting population. The population is constructed so that, wherever possible, any two structures were maximally different from each other and contained different helices.

3. Calculation of the energy of all structures. Dwell on this step, because its efficiency is one of the decisive factors for the general speed of the algorithm.

The energy of an RNA SS is calculated by a rapid recursive procedure. The structure is represented as a binary tree (Fig. 1). The nodes of the tree correspond to helices, and the edges connecting the nodes, to all loops except for hairpin ones. The helix most proximal to the 5'-end of the sequence (helix 1 in Fig. 1) is the root of the tree, and terminal leaves of the tree correspond to hairpins. Each node has two pointers: pointer *a* to a subtree (substructure), which is closed with the helix corresponding to the node, and pointer *b* to the substructure next to this helix in the 3'-direction. The energy of this structure is calculated by ordered visitation of each node of the tree. The sequence of steps for each node is as follows: first according to pointer *a*, then to pointer *b*, and return to the upper level. This visitation is called "from top to bottom" and allows a simple recursion (Wirth, 1986). In the example shown in Fig. 1, the nodes of the tree will be visited in the following order: 123456. Each step is accompanied by addition of the energy of the helix corresponding to the node visited and the energy of the passed loop to the energy of the structure. The thermodynamic rules from (Jaeger et al., 1989) are used.

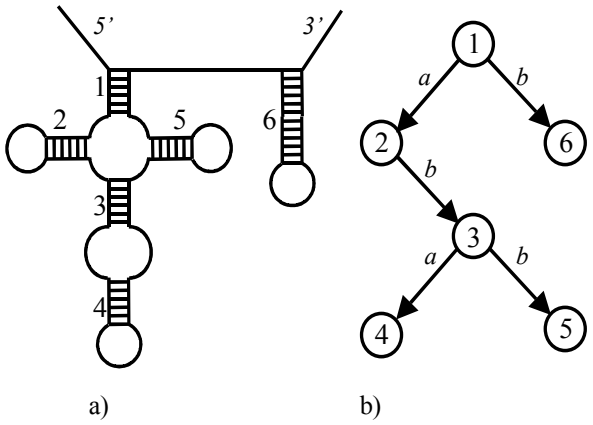


Fig. 1. (a) RNA secondary structure and (b) it's representation in the form of a binary tree.

1. Selection. The decision of the survival of a structure is made on the base of a random procedure in which the probabilities of survival are determined by the difference between the adaptability of the individual and the value averaged over the population. The adaptability of an individual (secondary structure) is calculated as follows: $f_i = \exp(-\frac{E_i}{\Delta E})$, where $E_i < 0$ is the free energy of the structure *i*, and $\Delta E > 0$ is the efficient resolution with regard to energy, i.e., such a difference in structure energies at which the ratio of their adaptability is equal to e.

2. Mutations. By a mutation a local change of RNA SS is meant. (a) A specified number of structures for mutation are randomly chosen from the population; (b) A specified number of helices are removed from each structure; (c) Each structure is sequentially supplemented with helices most favorable with regard to the structure stability; (d) If the resulting structure ranks below the initial structure in energy, the result of the mutations is discarded.

3. Recombinations: In the task under consideration, we mean by a recombination an exchange of large SS blocks. In our algorithm, recombinations are aimed at an equal and, therefore, the greatest difference between the descendant and its parents. This ensures the most large-scale search. Two randomly chosen parent structures and their common descendants form the framework of the descendant structure. Then the framework is supplemented with helices from the common list {h}, provided that the addition of a helix decreases the free energy of the structure.

4. Calculation stop: Steps 3–6 are repeated until a strong degeneracy of the population (similarity of all structures) is achieved. The similarity between two structures is estimated as the ratio between the number of matching base pairs and the total number of pairs in both structures. The calculation is halted when the structure similarity averaged over the population exceeds a certain limit.

Our calculations showed that the complexity of the algorithm at lengths below 400 nucleotides equals $O(N^{2.5})$, which matches the complexity of the dynamic algorithm mfold (Mathews et al., 1999) to a proportionality factor (Fig. 2).

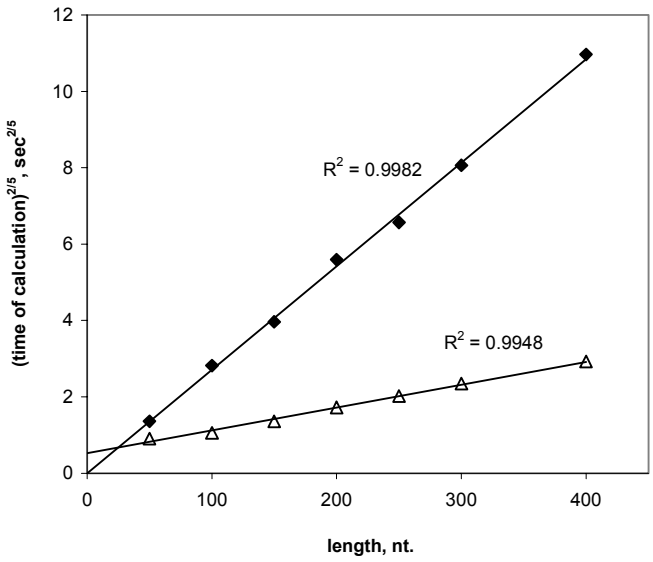


Fig. 2. Relationship between the time of culculation of the GA secondary structure (rhomb) and the algorithm mfold (triangles) and the sequence lenght (means for 50 random sequences of equal composition).

II. Blocks for calculating Z-score and E-score

We enlarge on a peculiar feature of our Internet resource: the opportunity to calculate the indices *Z score* and *E-score.* They are defined as

$$Z(L = const) = \frac{E - \langle E_i^{rand} \rangle}{\sqrt{disp(E_i^{rand})}},$$

where E is the structure energy of a natural sequence, E_i^{rand} is the structure energy for a random sequence formed from a natural one by nucleotide mixing, and

E-score = $9n_gn_c + 3n_an_u + 2n_gn_u$,

where $n_{g,c,u,a}<1$ is the frequency of a nucleotide in the sequence. The coefficients before the pairs of frequencies of complementary nucleotides G-C, A-U, and G-U reflect their energy contribution during SS formation.

Earlier, we showed that the E-score value calculated for random RNA sequences correlate well ($r^2 = 0.89$) with the energies of their SSs (Titov et al., 2002).

Also, we showed that for a sample of random sequences of variable lengths and compositions the distribution of Z-score is close to normal one with the mean value equal to zero and the variance equal to unity (Titov et al., 2002). Therefore, Z-score is suitable for comparing the stability of an RNA SS in samples heterogeneous in length and composition with the use of conventional statistical tests.

Hence, E-score reflects the contribution of nucleotide composition into SS stability (effect of composition). The index Z-score illustrates the indistinctiveness of SS stability (i.e. uses E-score as the null hypothesis and reflects the degree of orderly arrangement of nucleotides).

Because of the requirement of mutation stability (Fontana et al., 1993; Bonhoeffer et al., 1993) and rapid and efficient RNA folding into the target structure, the nucleotide composition of structural RNAs, unlike other genomic sequences, ranges within narrow limits (Titov et al., 2002). Therefore, E-score can be used for search for RNA in a genome.

A negative Z-score value points to an elevated stability (and, as a rule, evolutionary selection) of an RNA SS. For example, the typical Z-score value for tRNAs and 5S-RNAs is -1.8 (Titov et al., 2002). Negative values can be expected for other structural RNAs, which also can be helpful for their search in a genome.

Implementation

The algorithm has been implemented in the C language of the ANSI standard and installed in Internet at the address http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna/. The Internet version of the algorithm can calculate the structures of RNA molecules up to 250 nucleotides in length. The analysis previously performed by us showed that calculations within this range of lengths practically always yield the best solution (Titov et al., 2002). Several runs of algorithm with different locations of the initial population ensure the finding of the optimum solution.

The WEB interface of the algorithm is designed as follows: The user enters a nucleotide sequence in the text format and some parameters of the algorithm in separate windows of the browser. The parameters are: (1) minimum stem length; (2) for modeling interaction with an oligonucleotide, numbers of nucleotides with which it is paired; (3) the threshold value of population degeneracy for halting the calculation; (4) selection temperature; and (5) randomization index. At the output, the user obtains: (1) the energy of the predicted SS; (2) Z-score of the energy; (3) SS in the text format readable by other packages for RNA SS calculation, including the mfold program (*.ct file); and (4) graphical presentation of the SS in the window of a Java application allowing the image to be moved or resized.

Discussion

The high calculation speed of our algorithm is achieved owing to (1) the rapid recursive calculation of energy; (2) precalculation of stem energies; (3) the most uniform covering of the stretch of possible secondary structures with the initial population; (4) controllable "mutagenesis" (survival of adaptive mutations only); and (5) symmetrical recombinations.

In addition to prediction of SSs for an RNA sequence, the resource allows calculation of E- and Z-scores, which allow conclusions of the evolutionary significance of the SS for the chosen sequence.

Together with comparative analysis methods and search for structural motifs (S- and U-turns, tetraloops, etc.) and methods for search for RNA polymerase III promoters, the indices E-score and Z-score are applicable to search for new RNAs in a genome.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grant № 01-07-90376); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors are grateful to V.V.Gulevich and G.B.Chirikova for assistance in translation.

References

1. Benedetti G., Morosetti S. (1995) A genetic algorithm to search for optimal and suboptimal RNA secondary structures. Biophys. Chem. 55, 253.
2. Bonhoeffer S., McCaskill J.S., Stadler P.F., Schuster P. (1993) RNA multi-structure landscapes. A study based on temperature-dependent partition functions. Eur. Biophys. J. 22, 13–24.
3. Currey K.M., Shapiro B.A. (1997) Secondary structure computer prediction of the poliovirus 5' non-coding region is improved by a genetic algorithm. Comput. Applic. Biosci. 13, 1.
4. Fontana W., Konnings D.A.M., Stadler P.F., Schuster P. (1993) Statistics of RNA secondary structure. Biopolymers. 33, 1389.
5. Gultyaev A.P., van Batenburg F.H.D., Pleij C.W.A. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. J. Mol. Biol. 250, 37–51.
6. Jaeger J.A., Turner D.H., Zuker M. (1989) Improved predictions of secondary structures for RNA. Proc. Natl Acad. Sci. USA. 86, 7706.
7. Mattews D.H., Sabina J., Zuker M., Turner D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940.
8. McCaskill J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 29, 1105–1119.
9. Mironov A.A., Dyakonova L.P., Kister A.E. (1985) A kinetic approach to the prediction of RNA secondary structures. J. Biomol. Struct. Dyn. 2, 953–962.
10. Proutski V., Gould E.A., Holmes E.C. (1997) Secondary structure of the 3' untranslated region of flaviviruses: similarities and differences. Nucl. Acids Res. 25, 1194–1202.
11. Wirth N. Algorithms and data structure. New Jersey, Prentice Hall, Inc., Englewood Cliffs, 1986.
12. Titov I.I., Vorobiev D.G., Ivanisenko V.A., Kolchanov N.A. A fast genetic algorithm for RNA secondary structure analysis. Chem. Bull., in press.

ALGORITHM FOR SEARCHING FOR ALTERNATIVE SECONDARY RNA STRUCTURES

* ***Lyubetsky E.V., Lyubetsky V.A.***

Institute for Information Transmission Problems, RAS, Moscow, Russia, e-mail: Lin@iitp.ru

Key words: *secondary structure hairpin, alternative structure, terminator, coordinates of hairpin loops, algorithm, statistical certainty*

Resume

Motivation: A new algorithm and a model of searching for alternative secondary RNA structures and their specific peculiarities in including hairpin loops were implemented.

Results: The algorithm has shown a good efficiency on more than 120 experimental natural RNA fragments.

Availability: The software is available on request directed to authors.

Introduction

The problem of prediction of alternative secondary RNA structures is considered. It was found out last years that such structures being considered at mRNA level play an unexpectedly important role taking part in regulation of biosynthesis processes in a cell (attenuation). In models of regulation, the essential role is allocated to just unpaired bases, in particular, to the nucleotides of hairpin loops. The problem arises to find a small number of candidates for the alternative secondary structure in a given fragment of an RNA sequence (or to declare that there is no such structure).

An algorithm is proposed that is apparently one of the first algorithms aimed to solve this problem. Therefore, we had no opportunity to compare its results with processing of other algorithms. The algorithm was tested on random sequences and also in situations when the alternative secondary structures were retrieved experimentally.

The algorithm was applied in rather general situation, but here we present only the results of search for "three hairpin" attenuation and T-box terminator–antiterminator structures as well as of search for coordinates of hairpin loops and other hairpin patterns in such structures.

The algorithm was tested on the following structures:

1) Transcription attenuators for pheA and trp genes (which are used in biosynthesis of aromatic amino acid in gamma-proteobacteria) and for pheS gene (coding for phenylalanine-tRNA synthetase). We tested 17 such structures: the results were near 100% (see the list the table below).

2) Transcription attenuators for pyrimidine biosynthesis genes in *B. subtilis*. We tested three such structures: loop coordinates of the secondary hairpin structure regulating them were retrieved (see rows 1–3 in table).

3) T-box terminator–antiterminator structure (taking part in regulation of genes pertaining to biosynthesis of amino acids and genes coding for aminoacyl-tRNA synthetase in gram-positive bacteria). Loop coordinates and terminator hairpin of the secondary hairpin structure regulating them were retrieved (see rows 4–43 in the table).

Let us remark that the algorithm does not use any specific parameter of the secondary structure; it was not supported by experimental answers anyway.

Methods and Algorithms

The algorithm is based on recursive operating with some set of hairpin and structure parameters and proceeds from parameters to hairpins and structures only at the last stage of the processing. The following hairpin parameters are used by the algorithm: power, the origin A and endpoint B of the left half-stem, the origin C and endpoint D of the right half-stem of a hairpin, free energy, and so on. A structure parameter linking power of two hairpins is also used. The algorithm generates a set of locally optimal parameters and used it to produce the set of locally optimal hairpins. The last of them is statistically analyzed to build a consensus hairpin and an alternative secondary structure from the obtained hairpins. A more detailed description of this rather logically intricate, but fast and efficient algorithm can be found in (Vereshchagin, Lyubetsky, 2000).

The algorithm was implemented as a program in Object Pascal language in Delphi 5 environment and also in ANSI C in a serial and parallel computing architectures. It showed a high efficiency on some natural mRNA sequences.

* Corresponding author.

Implementation and Results

Some results of the algorithm testing is shown in table and in the list following it. Here, Sh is a specifier hairpin, A is an antiterminator, and T is a terminator. Expression "exact" in the 4th column means that loops of hairpins from one alternative secondary structure are found in exact correspondence with experimental answer. "NF" (not found) means that the coordinates of hairpin loops differ from the experimental answer on more than 10 positions in the hairpin left or right half-stems. Expressions like "8,10" mean that difference between (B,C) values in algorithm results and experimental answers is 8 nucleotides in the left half-stem and 10 nucleotides in the right one for a hairpin under consideration. Expression like "W2" (without 2 pairs) in the 5th column means that at the end of the terminator hairpin, located by the algorithm, 2 nucleotide pairs from experimental answer are missing.

	Gene name	Number of biol. hairpes	Precision of locating of hairpin loop coordinates	Precision of term-or locating
1	Bs_pyrB	2	T – exact; A – exact	---
2	Bs_pyrP	2	T – 8,10; A – 4,1	---
3	Bs_pyrR	2	T – exact; A – exact	---
4	Be_serS	5	T, Sh, 2, 3, A – exact	Exact
5	Be_tyrS	5	T – exact; Sh - 2,5; 2 – NF; 3, A – exact	Exact
6	Bq_serS	5	T – exact; Sh – NF; 2 – 2,3;3 – exact; A – 0,2	Exact
7	Bq_tyrS1	5	T – exact; Sh – exact; 2,3 – NF; A – exact	NF
8	Bq_tyrS2	5	T – exact; Sh – exact; 2 – 0, 3; 3, A – exact	W3
9	Bs_serS	5	T – 1,1; Sh – 3,5; 2,3,A – exact	W1
10	Bs_tyrS	6	T – exact; Sh – NF; 2 – 2,2; 3 – 3,3; 4 – exact; A – 0,1	Exact
11	Bs_tyrZ	6	T – NF; Sh – exact; 2 – NF; 3 – 1,0; 4 – exact; A – NF	NF
12	Ca_tyrZ	3	T – NF; Sh – 1,4; A – 0,1	NF
13	Ca_yurG	5	T – 0,1; Sh – 4,1; 2,3 – NF; A – 2,0	NF
14	DF_serS	5	T – NF; Sh, 2, 3 – exact; A – 4,1	NF
15	DF_tyrZ	3	T – 1,0; Sh – 4,1; A – exact	W1
16	DHA_tyrZ	3	T – 0,1; Sh – 0, 2; A – 0,1	NF
17	EF_serS	3	T – 3,2; Sh – NF; A – 1,4	NF
18	EF_tyrS	5	T – exact; Sh – NF; 2 – exact; 3 – 6,2; A – exact	NF
19	HD_serS	5	T – exact; Sh – exact; 2 – 0,6; 3 – 5,3; A – 0,1	NF
20	HD_tyrZ	6	T – exact; Sh – 5,3; 2 – NF; 3 – 4,1; 4 – 1,1; A – exact	NF
21	LLX_serS	3	T – 1,2; Sh – 1,0; A – 0,1	NF
22	LO_serS	3	T – 0,1; Sh – 3,1; A – NF	Part. found
23	LO_tyrS	5	T – exact; Sh – 4,1; 2,3 – exact; A – 0,1	Exact
24	PN_serS	3	T – 1,2; Sh – 2,6; A – exact	NF
25	Sa_serS	5	T – exact; Sh – 5,3; 2 – 0,1; 3,A – NF	Exact
26	SEQ_serS	3	T – 1,1; Sh – NF; A – exact	NF
27	Bs_thrS	5	T – 1,0; Sh – NF; 2,3,A – exact	NF
28	LL_his	5	T – 5,4; Sh,2,3 – exact; A - NF	NF
29	LL_trp	5	T – exact; Sh – exact; 2 – NF; 3– 0,2; A – exact	Exact
30	Bs_purE	3	T – NF; Sh – exact; A – NF	NF
31	Bs_purM	1	T – 2,1	W2
32	Bs_tyrS	5	T – exact; Sh, 2 – NF; 3, A – exact	Exact
33	Ec_pyrB	2	T – 0,1; A – NF	NF
34	Ec_ilvG	3	T – exact; Sh, A – exact	NF
35	Ec_rpsJ	6	1 – NF; 2 – 1,1; 3,4 – NF; 5 - exact; 6 – 3,2	No term-r
36	Hi_rpsJ	5	1 – 3,1; 2 – 1,1; 3,4 – exact; 5 - 6,2	No term-r
37	Bs_ilv_leu	5	T – exact; Sh – exact; 2 – 2,6; 3 – NF; A – exact	NF
38	Bs_yczA	5	T – exact; Sh – NF; 2 – 1,1; 3 – exact; A – NF	Exact
39	Bs_trpE	2	1 – exact; 2 – 0,1	No term-r
40	Sa_ileS	4	T – exact; Sh, 2, A – NF	NF
41	Ec_rplK	2	1, 2 – exact	No term-r
42	Hi_rplK	2	1, 2 – exact	No term-r
43	Bs_valS	3	T – 0,1; Sh – 4,3; A – NF	NF

In the following list are names of genes followed by evaluations of likeness of alternative secondary structures found experimentally and by the algorithm. The first one is evaluation of the terminator hairpin; the second and third are evaluations of antiterminator and specifier hairpins of alternative secondary structures. Evaluation "2" means here that the distances between two answers on B and on C are less then 5, and on A and on D are less then 7. Similarly evaluation "1" means that these distances are less then 5 and more then 7. Finally, evaluation "0" means that these distances are more then 5 and more then 7. So, we have the following comparisons of the algorithm results with the experimental answers: Aa_aroma_pheA 210, 000; Ec_aroma_pheA 000, 000, 121; Ec_aroma_pheS 221, 000; Ec_aroma_trpE 210, 000; Hi_aroma_pheA 210; Hi_aroma_pheST 210; Hi_aroma_trpBA 000; Hi_aroma_trpE 000, 000; St_aroma_pheA 210, 000, 000; St_aroma_pheS 212, 000; St_aroma_trpE 210; Vc_aroma_pheA 211; Vc_aroma_trpE 210, 000, 000, 000; Yp_aroma_pheA1 221; Yp_aroma_pheA2 210; Yp_aroma_PheS 000, 211, 000, 000; and Yp_aroma_trpE 210. Let us remark that in three cases including two of them having incorrect answers, terminator hairpins are really more complicated (they include bulges). Now, we have developed an algorithm for locating such terminator hairpins.

Discussion

1) Loop coordinates of many hairpins (e.g. terminator hairpins) are located with a very high precision. Terminator hairpins of "three hairpin" attenuation structures are located in 15 cases of 17 (88%). Terminators are located in 29 cases (56%) of 52.

2) When testing 129 sequences, we found that the number of hairpin loop coordinates repeated more then 9 times, was rather small relative to the total number of hairpin loop coordinates that were located by the algorithm. The total number of locally optimal hairpins for each studied sequence was 300–600, the algorithm has selected only 10–25 loci in it for further processing.

3) Hairpin loop coordinates and other hairpin and configuration patterns found by the algorithm are almost independent of their locations in the original sequence and its length.

4) In most cases when the algorithm found an answer with a low precision, the biological answer was also "incorrect" in the sense that hairpins contained side subhairpins or sections of power 2 or pairs <G,T> at the endpoints of sections. An algorithm is now developed that locates such hairpins with the same efficiency.

5) In the case of exact retrieving of hairpin loop coordinates, hairpin's first section was located almost always correctly. In half of the cases, two sections were located correctly. In some cases, whole hairpin was retrieved (all its 1–5 sections).

The detailed results of research are placed in the online journal *Information Processes* at http://www.jip.ru.

Acknowledgements

The authors express a deep gratitude to Prof. M.S.Gelfand and A.A.Mironov for their help and explanations of the biological content of the problem. Computer program was implemented and run by L.A.Leontiev.

References

1. Vereshchagin N.K., Lyubetsky V.A. (2000) Algorithm for determination of alternative secondary RNA structures. Transact. Research Seminar of the Logical Center of the Institute of Philosophy RAS, 14, M.: Nauka, 99-109.

ALGORITHM FOR PREDICTING THE EVOLUTIONARILY CONSERVED SECONDARY STRUCTURES OF RNA

Vorobiev D.G.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: denis@bionet.nsc.ru

Key words: *RNA, secondary structure, genetic algorithm, phylogenetic approach*

Resume

Motivation: Phylogenetic algorithms for predicting the RNA secondary structure (SS) offer advantage over thermodynamic methods in the presence of representative samples of isofunctional sequences of RNA.

Results: In this paper, a new phylogenetic algorithm for constructing the RNA SS from a multiple alignment of sequences using a genetic algorithm approach is presented.

Availability: the program is available from the author.

Introduction

The RNA SS is predicted using approaches referring to two groups: thermodynamic and phylogenetic algorithms. The former are most extensively used (Mathews et al., 1999; McCaskill, 1990; Gultyaev et al., 1995) because their operation needs no other data than RNA sequences. The experimental evidence for either the paired or free states of separate bases simplifies calculation of the RNA secondary structure. However, these data are not always available. The phylogenetic algorithms are used more rarely because they need a consistent sampling of RNA sequences with a similar function and structure.

The errors of thermodynamic algorithms usually occur for one of three reasons: the inaccuracy of energy rules, the impossibility of taking into account the energetics of either tertiary or RNA-protein interactions and the disregard of the peculiarities of folding kinetics. The last problem was partilly solved by the method of RNA ensemble kinetics modelling (see, Mironov et al., 1986).

The phylogenetic algorithms use not only RNA thermodynamics but also information on the phylogenetic conservatism of RNA SS that perform a similar function (Eddy, Durbin, 1994; Gorodkin et al., 1997; Chen et al., 2000; Hofacker, Stadler, 1999). This offers them a fundamental advantage and in many cases, allows them to get around the drawbacks typical of purely thermodynamic methods.

In this paper, a pilot variant of the algorithm for predicting the evolutionarily stable RNA is proposed. Actually, it is similar to genetic algorithm (GA) used to predict the SS of a single RNA molecule (Vorobiev et al., this volume). However, it allows one to take into account the conservatism of the primary (high homology of sample sequences) and secondary (coadaptive substitutions) structures.

Methods and Algorithms

Assume that we have a multiple alignment of N sequences of RNA of length L (problems on the construction of such alignment will be discussed below). The RNA SS is unambiguously given by a set of helices. In the unit RNA sequence, the helix is represented by the three <x,y,l> (where x is the coordinate of the left end of the helix arm, y is that of the right end of the helix arm and l is the helix length) such that the bases at the sequence positions x+a and y-a (for all $0 \le a < 1$) form the complementary pairs (AU, GC, or GU).

In the first step of the algorithm, we make up the list {h} of helices common for the sequences from the alignment. The helix is included in the list {h} if a continuous (without gaps) helix with the same coordinates is present in more than T_{share}% of sequences.

The value of T_{share} is set by the user. The energy of such a helix is defined by us as the sum of helix energies in sequences, having the helix with the same coordinates, devided by N. We used the thermodynamic parameters from (Jaeger et al., 1989). The term $P_c n_c$ can be added to the energy of helix x,y,l if it is contained in all sample sequences. In this case, P_c is the prize set by the user for a pair of positions x+a, y-a ($0 \le a < 1$) where nucleotides in the different sequences of the alignment form different pairs ($n_c \le 1$ being the number of such pairs of positions in the given helix). The summand $P_c n_c$ of energy should take into account the coadaptive substitutions.

In the second step, the list {h} is send to the input of the genetic algorithm identical to that used by us to predict the SS of the single RNA molecule (Vorobiev et al., this volume). On the output of the genetic algorithm there is a set of helices some of which cannot exist in all sequences. For these sequences, the forbidden helices are broken either partially or completely.

For each sequence, we can optionally complete the construction of helices, reducing the energy of its SS, by the steepest descent technique.

Implementation and Results

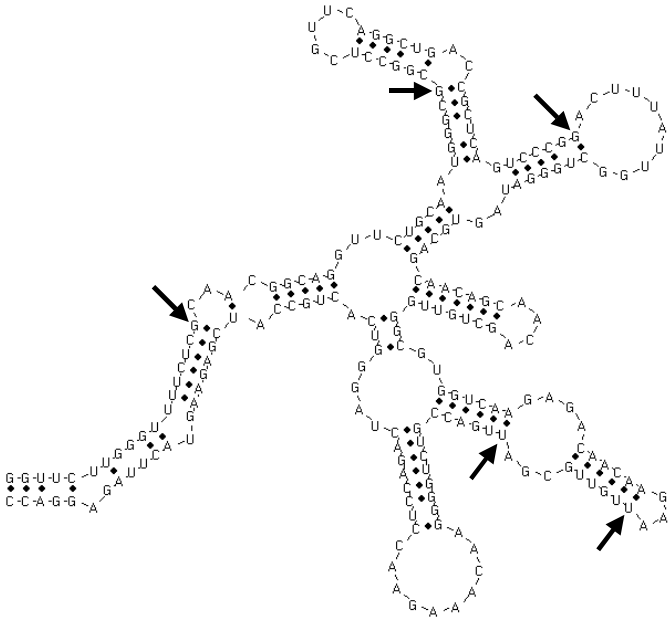
The algorithm has been realized in language C of standard ANSI. It was tested using RNA of HIV-2 rev response elements (Fig.) and 5S RNA (Chen et al., 2000). In these cases, more than 90% of nucleotide pairs forming a real structure were predicted.

```
         SIVMM251
GGUUCUUGGGUUUUCUCGCAACGGCAGGUUCUGCAAUGGGCGCGGCCUCGUUCAGGCUGA
         SIVMM142
GGUUCUUGGGUUUUCUCGCAACGGCAGGUUCUGCAAUGGGCGCGGCGUCGUUGACCGUGA
         HIV2ROD
GGUUCUUGGGUUUUCUCGCAACAGCAGGUUCUGCAAUGGGCGCGGCGUCCCUGACCGUGU
         HIV2BEN
GGUUCUUGGGUUUUCUCGCGACAGCAGGUUCUGCAAUGGGCGCGCGGUCCCUGACGCUGU
         HIV2CAM2
GGUUCUUGGGUUUUCUCACAACAGCAGGAGUUGCAAUGGGCACGGCGUCCUUGACGCUGU
         HIV2NIHZ
GGUUCCUAGGUUUUCUCGCAACAGCAGGUUCUGCCAUGGGCGCGGCGUCCUUGACGCUGU
         HIV2ISY
GGUUCCUAGGUUUUCUCACGACAGCAGGUGCUGCAAUGGGGGCGGCGUCUCUGACGCUGU
         HIV2ST
GGUUCUUAGGUUUUCUCACGACAGCAGGAGCUGCAAUGGGCGCGGCGUCCUUGACGCUGU
         HIV2D194
GGUUCUUGGGUUUUCUCGCGACAGCAGGUUCUGCAAUGGGCGGCGCGUCCUUGACGCUGU
         HIV2UC1
GGUUCUUGGGACUUCUUGCAAUGGCAGGUUCUGCAAUGGGCGCAACGUCCUUGACGCUGU

         SIVMM251
CCGCUCAGUCCCGGACUUUAUUGGCUGGGAUAGUGCAGCAACAGCAACAGCUGUUGGGCG
         SIVMM142
CCGCUCAGUCCCGGACUUUAUUGGCUGGGAUAGUGCAGCAACAGCAACAGCUGUUGGACG
         HIV2ROD
CGGCUCAGUCCCGGACUUUACUGGCCGGGAUAGUGCAGCAACAGCAACAGCUGUUGGACG
         HIV2BEN
CAGCCCAGUCCCGGACUUUACUGGCCGGGAUAGUGCAGCAACAGCAACAGCUGUUGGACG
         HIV2CAM2
CAGCCCAGUCUCGGACUUUAUUGGCCGGGAUAGUGCAGCAACAGCAACAGCUGUUAGACG
         HIV2NIHZ
CAGCUCAGUCUCGGACUUUAUUGGCCGGGAUAGUGCAGCAACAGCAACAGCUGUUGGAUG
         HIV2ISY
CGGCUCAGUCUCGGACUUUAUUCCGUGGGAUAGUGCAGCAACAGCAACAGCUGUUGGACG
         HIV2ST
CGGCUCAGUCUCGGACUUUAUUGGCCGGGAUAGUGCAGCAACAGCAACAGCUGUUGGACG
         HIV2D194
CGGCUCAGUCCCGGACUUUACUGGCCGGGAUAGUGCAGCAACAGCAACAGCUGUUGGACG
         HIV2UC1
CAGCUCAGUCCCGGACUUUACUGGCUGGGAUAGUGCAGCAGCAACAGCUGCUGGACG

         SIVMM251
UGGUCAAGAGACAACAAGAAUUGUUGCGAUUGACCGUCUGGGGAACAAAGAACCUCCAGA
         SIVMM142
UGGUCAAGAGACAACAAGAAUUGUUGCGACUGACCGUCUGGGGAACAAAGAACCUCCAGA
         HIV2ROD
UGGUCAAGAGACAACAAGAACUGUUGCGACUGACCGUCUGGGGAACGAAAAACCUCCAGG
         HIV2BEN
UAGUCAAGAGACAACAAGAAAUGUUGCGACUGACCGUCUGGGGAACGAAAAACCUCCAGG
         HIV2CAM2
UGGUCAAGAGACAACAAGAAUUGUUGCGACUGACCGUCUGGGGAACAAAAAUUCUCCAGG
         HIV2NIHZ
UGGUCAAGAGACAACAAGAAAUGUUGCGACUGACCGUCUGGGGAACAAAAAAUCUCCAGG
         HIV2ISY
UGGUCAAGAGACAACAAGAAAUGUUGCGACUGACCGUCUGGGGAACUAAAAACCUCCAGG
         HIV2ST
UGGUCAAGAGACAACAAGAAAUGUUGCGACUGACCGUCUGGGGAACAAAAAAUCUCCAGG
         HIV2D194
UGGUCAAGAGACAACAAGAAAUGUUGCGAUUGACCGUCUGGGGAACGAAAAAUCUCCAGG
         HIV2UC1
UGGUCAAAAGACAACAGGAACUGUUGCGGCUGACCGUCUGGGGAACGAAAAACCUCCAGA

         SIVMM251 CUAGGGUCACUGCCAUCGAGAAGUACUUAGAGGACC
         SIVMM142 CUAGGGUCUCUGCCAUCGAGAAGUACUUAAAGGACC
         HIV2ROD  CAAGAGUCACUGCUAUAGAGAAGUACCUACAGGACC
HIV2BEN  CAAGAGUCACUGCUAUCGAGAAGUACCUAAAGCAUC
         HIV2CAM2 CAAGAGUCACUGCUAUAGAGAAGUACCUAAAGGAUC
         HIV2NIHZ CAAGAGUCACUGCUAUAGAGAAGUCACUAAAGGACC
         HIV2ISY  CAAGAGUCACUGCUAUUGAGAAGUACCUAGCAGACC
         HIV2ST   CAAGAGUCACUGCUAUCGAGAAAUACUUAAAGGACC
         HIV2D194 CAAGAGUCACUGCUAUCGAGAAAUACUUAAAGGACC
         HIV2UC1  CAAGAGUCACUGCCAUCGAGAAAUACCUAAAGGACC
```



a) **b)**

Fig. An example of algorithm operation: (a) the multiple alignment of the RNA sequence of HIV rev response elements send to the program input, and (b) the structure obtained for the first sequence from the alignment. The arrrows denote the pairs added using the method of steepest descent to the structure "core" arising at the GA output.

Discussion

The main problem of our approach as of all other similar algorithms is the need for preliminary construction of a multiple alignment of RNA sequences. This restricts the range of approach applicability to the cases of high homology of RNA sequences. The algorithms like ClustalW (Thompson et al., 1994), usually used to construct the alignment, are adequate to the problem of SS consensus construction only in the case of high sequence homology in a sample. Therefore, it is of prime importance that our algorithm is supplemented with a method for constrocing multiple alignment that takes into account the conservatism by both the primary and secondary structures.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (grant № 01-07-90376), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), National Institutes of Health USA (№ 2 Ro1-HG-01539-04A2), the Department of Energy USA (№ 535228 CFDA 81.049). The author is grateful to N.A.Kolchanov and I.I.Titov for helpful discussions and to G.A.Ilyina for assistance in translation.

References

1. Chen J.H., Le S.-Y., Maizel J.V. (2000) Prediction of common secondary structure of RNAs: a genetic algorithm approach. Nucl. Acids Res. 28**,** 991-999.
2. Gultyaev A.P., van Batenburg F.H.D., Pleij C.W.A. (1995) The computer simulation of RNA folding pathways using a genetic algorithm. J. Mol. Biol. 250**,** 37-51.
3. Hofacker I.L., Stadler P.F. (1999) Automatic detection of conserved base pairing patterns in RNA virus genomes. Comput. Chem. 23, 401-414.
4. Jaeger J.A., Turner D.H., Zuker M. (1989) Improved predictions of secondary structures for RNA. Proc. Natl Acad. Sci. USA. 86**,** 7706.
5. Mathews D.H., Sabina J., Zuker M., Turner D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911-40.
6. McCaskill J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 29, 1105-1119.
7. Mironov A.A., Dyakonova L.P., Kister A.E. (1985) A kinetic approach to the prediction of RNA secondary structures. J. Biol. Struct. Dyn. 2**,** 953-62.
8. Thompson J.D., Higgs D.G., Gibson T.J. (1994) CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalies, and eight matrix choice. Nucl. Acids Res. 22, 4673-4680.
9. Vorobiev D.G., Titov I.I., Ivanisenko V.A. (2002) GArna Internet resource for the analysis of the RNA secondary structure:its status in 2002. This volume.

AN ALGORITHM FOR SEARCHING FOR COMMON SECONDARY STRUCTURES IN A SET OF RNA SEQUENCES

* ***Gorbunov K.Yu.***, ***Lyubetsky V.A.***

Institute for Information Transmission Problems RAS, Moscow, Russia, e-mail: lyubetsk@iitp.ru, gorbunov@iitp.ru

Key words: *secondary RNA structure, common structure, alignment, tRNA*

Resume

Motivation: We propose an algorithm for searching for conservative secondary structures in a set of RNA sequences. Its complexity is quadratic in the sum of lengths of the input sequences. The main idea of the algorithm is a concurrent alignment of sequences of possible structure elements.

Results: The algorithm was tested on various kinds of conservative secondary RNA structures. Practical applicability of the algorithm was demonstrated—about 70–80% of the biological hairpins were found by this algorithm.

Availability: The software is available on request directed to authors.

Introduction

Regulatory RNA secondary structures are often similar in related genomes. This raises the problem of predicting such structures in a family of RNA sequences. This problem seems to be far from effective algorithmic solution in general situation. Secondary structure consists of hairpins. Every hairpin consists of two ordered sequences of segments located from left to right at each sequence having some bulges between neighbor segments. Under hairpin loop, we mean the segment between these ordered sequences, which are called half-stems of the hairpin. Each *i*th segment from the beginning of the left half-stem serves as a complement of the *i*th segment from the end of the right half-stem. The pair of such *i*th segments is called a helix. In other words, helix is an equivalent of a hairpin having only one segment in each sequence. Thus, we can consider helices as elementary parts of a secondary structure. Often, secondary structure contains quite long helices.

The known algorithms for prediction of secondary structures are based on comparative retrieval of the corresponding structures in a given family of RNA sequences. For example, the method of dynamic programming is used (Gorodkin et al., 1997) to construct secondary structures that are both similar and maximally powerful for every pair of sequences and for every pair of their subsequences (starting with short subsequences). An inference of structures in stochastic context-free grammars is used with the same object (Eddy, Durbin, 1994), and so on.

Our algorithm is based on a different approach. We start from a representation of secondary structures as a linearly ordered set of (left and right) half-stems of hairpins, not as a tree of hairpins. The hairpins are placed in ascending order of their coordinates. The coordinate of a left half-stem is defined as the number of position of the rightmost nucleotide in it. The coordinate of a right half-stem is defined as the number of position of the leftmost nucleotide in it. The half-stems with the same coordinates are ordered arbitrarily. In conservative structures, homologies of half-stems are ordered similarly. This demonstrates the main idea of our algorithm—a concurrent alignment of half-stems of hairpins.

Methods and Algorithms

For any n given RNA sequences, we construct a large list of possible helices with a length not less than a prescribed value and with the distance between half-stems lying in a prescribed range (we construct only the helices continuing in both directions as far as possible taking into account that helices often have this property in secondary structures). We join into hairpins those helices whose left and, respectively, right half-stems are located at the small distance from each other. So, n lists of hairpins $L1,\ldots, Ln$ will be obtained.

For each pair of hairpins from different lists, we estimate their similarity. Thus, the base of similarities of hairpins is created. In biological hairpins, similar regions can lie in both helices and bulges, or even outside of a hairpin at a small distance from it. Our program enables us to take into account these possibilities: there are parameters defining the way of constructing a word from a hairpin. The words constructed above are compared by the Smith–Waterman method (see, for example, Waterman, 1989). To be more precise, a variant of this method aimed to find the most similar subwords in given words was

* Corresponding author

used (we also used other algorithms of the same type). We can correct similarities estimated by this method by setting certain parameters: penalty for difference in length of hairpin loops, penalty for long hairpin loops, and so on. All the similarities that are less than a chosen threshold *t* are ignored (they are replaced with 0).

Now, our aim is to refine the lists *L*1,…, *Ln* so that they would contain only the hairpins that form the desired structure. At the first (rough) stage, we delete each hairpin *h* from each list so that that the number of lists containing a hairpin *h1* with similarity (*h,h1*)≥ *t* is small.

The second (main) stage of the refinement demonstrates the principal idea of the algorithm. We transform each list of hairpins into a list of their left and right half-stems ordered according to increase in their location (coordinate) in RNA sequence (for a left half-stem, we choose its end as coordinate; for a right half-stem, we choose its beginning as coordinate). We consider a list of half-stems as a word whose letters are half-stems. Thus, each pair of lists can be aligned with the above-mentioned variant of the Smith–Waterman algorithm (or by any other algorithm of that type). It is natural that we allow the left half-stems to match only with left half-stems and, analogously, for right half-stems. When two half-stems are matched, we take their similarity from the above-mentioned base—this is the similarity between the corresponding hairpins. Taking into account that, as a rule, many superfluous hairpins are present in our lists, it is reasonable to choose null or small penalty for deletion of a half-stem.

After each pair of lists has been aligned, we count the quality of every hairpin—a value that reflects how often its half-stems were matching. While counting the quality, we assign a special price for *complete matching* of the hairpin. The complete matching occurs when half-stems of a hairpin is matched with half-stems of the same hairpin. Some price is also assigned to a hairpin *h* when two hairpins being completely matching with *h* are completely matching to each other.

Hairpins with null quality are deleted from the lists. The remaining hairpins are involved in the second iteration of alignments, after which the qualities of the hairpins are calculated again. The second iteration proceeds similarly to the first iteration but with two differences. First, the similarity of two half-stems is not merely taken from the base but is updated with regard to the qualities ascribed to the corresponding hairpins at the first iteration. Second, calculation of the qualities after the second iteration is more rigorous: we take into account only the complete matching. The computer program allows any chosen number of iterations to be performed; however, testing showed that two iterations suffice as a rule.

After the second stage of refinement of the lists, we form a joined (for all the sequences) list *L* of hairpins in descending order of their qualities (we can bound the length of *L*). The last stage of the algorithm is constructing of secondary structures. In each sequence, a structure is built independently of other sequences by our modification of Nussinov–Jacobson algorithm. It is known (Nussinov, Jacobson, 1980) that this algorithm constructs the most powerful structure on a given sequence (and on its each subsequence) by the method of dynamic programming (starting with short subsequences). Our modification of this algorithm consists in the following. First, we use half-stems of hairpins of the list *L,* not nucleotides, as primary elements. Second, instead of the most powerful structure, we build the structure with a maximal sum of qualities of the hairpins.

Let us seek a structure of "clover leaf" kind, that is, one helix with a long loop containing several helices with short loops (for example, as in the structure of tRNA) within this loop. Our algorithm can be amplified as follows (analogous possibility is provided for other kinds of secondary structures). At the last stage of the algorithm, we allow pairing of hairpin half-stems only if this hairpin has short loop or has long loop with desired number of helices in it already constructed. Similar improvement is provided for the stage of half-stem alignment. Certainly, particular conserved nucleotides can also be taken into account.

Implementation and Results

Let us describe the result of testing of the algorithm on 18 fragments of *Escherichia coli* tRNA. Below, after the number of organism and the anticodon, we cite the helices of real (biological) structures that were found by the algorithm (i.e. that are present in the structure suggested by the algorithm). The letter H denotes the lower helix (handle); L, the left helix; U, the upper helix; and R, the right helix. The number in brackets indicates how many superfluous helices were output (absent in the real structures). Sometimes when a false helix F lies near a real helix H, the algorithm may output F instead of H. The results below contain one such sample; the distances between the left and right ends of the hairpin loops of the false helix H and the real helix R are indicated in square brackets.

DA1660 TGC: H,L,U,R(1); DA1661 GGC: H,L,U,R(1); DC1660 GCA: H,U,R(0);

DD1660 GTC: H,U,R(1); DE1660 TTC: H,R(2); DF1660 GAA: H,L,U,R(0);

DG1660 TCC: H,U,R(1); DG1661 GCC: H,L,U,R(1); DG1662 CCC: H,L,U,R(0);

DH1660 GTG: H,L,U,R[1,2](1); DI1660 GAT: H,L,U,R(0); DI1661CAT: H,L,U,R(1);

DK1660 TTT: H,L,U,R(0); DL1660 CAG: U,R(2); DL1661 TAG: H,R(2);

DL1662 CAA: H,U,R(2); DL1663 GAG: H,U,R(1); DL1664 TAA: H,U,R(0).

We also carried out an extensive testing of the algorithm for other kinds of regulatory secondary RNA structures including RFN structures, regulating riboflavin biosynthesis and transport genes in various bacteria (Vitreschak et al, 2002). The

detailed results of this testing are submitted for publication in the electronic Journal *Information Processes* (http://www.jip.ru).

Discussion

The program admits one more stage: comparison of the structures constructed with each other and indication of the consensus structure together with its (partial) maps for the given structures. Such maps provide a possibility to predict the hairpins of real structures that for some reasons were not found by the algorithm.

Let us remark that all the stages of this algorithm except for the last stage can work even in the case when a real structure contains pseudoknots, that is, hairpins containing only one half-stem of another hairpin in their loop. It seems natural to use at the last stage of our algorithm a recently suggested algorithm of Rivas & Eddy (1999) with the corresponding modifications; this algorithm is designed for the same purpose as Nussinov–Jacobson algorithm but admit existence of pseudoknots. Though a time bound of this algorithm in the worst case is quite high (sixth power) but due to the fact that a few hairpins usually remain for the last stage, the algorithm of Rivas & Eddy (1999) works fast (as our computations showed).

Acknowledgments

The authors thank M.S.Gelfand and A.A.Mironov for help and for numerous explanations of biological content of the problem.

References

1. Eddy S., Durbin R. (1994). RNA sequence analysis using covariance models. Nucl. Acids Res. 22:2079–2088.
2. Gorodkin J., Heyer L.J., Stormo G.D. (1997). Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucl. Acids Res. 25:3724-3732.
3. Nussinov R., Jacobson A.B. (1980). Fast algorithm for predicting the secondary structure of single-stranded RNA. Proc. Natl Acad. Sci. USA. 77:6309–6313.
4. Rivas E., Eddy S.R. (1999). A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol. 285:2053–2068.
5. Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. (2002). Regulation of riboflavin biosynthesis and transport genes by a conserved RNA structural element. This volume.
6. Mathematical Methods for DNA Sequences. Waterman M.S. (ed), CRC Press, Inc., Boca Raton, Florida, 1989 (translation into Russian, M.: Mir, 1999).

REVEALING AND FUNCTIONAL ANALYSIS OF tRNA-LIKE SEQUENCES IN VARIOUS GENOMES

Frenkel F.E., * Korotkov E.V.

Center "Bioengineering" RAS, Moscow, Russia, e-mail: katrin2@beingi.ac.ru
*Corresponding author

Key words: *tRNA-like sequences, functional analysis, repeat, MIR, LINE, SINE, computer analysis*

Resume

Motivation: As tRNA is a very ancient genetic structure it is significant to analyze its functional and evolutionary role. Though tRNA-like sequences earlier was found and described in many gene regions and most known repeats contain tRNA part (Mans RM, Pleij CW, 1991) there are still a lot of unknown entries. tRNA-like sequences also could help in revealing new genetic repeat families.

Results: Appearance of tRNA-like sequences was revealed in various genetic structures. Relative distribution through them is shown.

Introduction

To reveal tRNA-like sequences we have collected them first from tRNA Web database of Dr. Mathias Sprinzl and Dr. K.S.Vassilenko (http://www.uni-bayreuth.de/departments/biochemie/sprinzl/trna/). Modified method of Dr. E.V.Korotkov (Korotkov, 2000) was used to reveal tRNA-like sequences that have been diverged from source tRNAs during the evolution process. We utilized one of the later GenBank versions as a source of genomic data.

Model

Method of enlarged DNA nucleotide sequence similarity was used to detect highly divergent tRNA copies. Model applied is based on sequence family (tRNAs families for each aminoacid) patterns taking into account possible insertions and deletions in its derivatives. A positional matrix of nucleotide frequencies is built from source multiple alignment of family members.

Smith-Waterman sequence alignment algorithm (Waterman, 1995) underlies the model applied. Primary method was described earlier in detail (Korotkov, 2000). Several modifications was made to the algorithm to make it able to detect multiple insertions and deletions, and to calculate sequence weight coefficients according to its representation in family (rarely appeared and more unique sequences get higher weight in its family) (Sibbald, Argos, 1990).

It is a method of a dynamic programming. Weight matrix for each position in source tRNA was calculated:

$$v(i,j) = f(i,j)\ ln\{f(i,j)/p(i)\} \quad (1)$$

where $f(i,j)$ is the frequency of the base i at sequence(source tRNAs) position j.

It shows us the likelihood that the given base is located at the given position. Smith-Waterman dynamic programming algorithm implies finding best alignment via filling the alignment weight matrix F as

$$F(i,j) = \max\{\max_{k=1,dmax}\{F(i-k,j) - v_d(1+\log(k))\};\ \max_{k=1,dmax}\{F(i,j-k) - v_d(1+\log(k))\};$$
$$F(i-1,j-1) + v(S(i),j);0.0\};$$
$$F(0,0) = 0.0; F(i,0) = F(0,0) - v_d(1+\log(i)); F(0,j) = F(0,0) - v_d(1+\log(j)) \quad (2)$$

where i is a position in the sequence S(tRNA), j is a position in the subsequence (window) of analyzed DNA, $dmax$ is maximal allowed deletions/insertions number, v_d is a deletion/insertion weight and $v(i,j)$ is previously defined (1) tRNA weight matrix.

After filling matrix F we are creating the optimal way from its maximum element to the first zero element. Weight of the found local alignment equals to a difference between corresponding last(maximal) and first(zero) F element.

Statistical significance of the alignment was calculated by comparing the calculated weight of alignment assumed as a global one against weights of tRNA global alignment along random sequences with the same nucleotide composition. Global alignment differs from local one in absence of zero member during filling alignment matrix (1).

Significance score Z equals:

$$Z = (W_S - M(W_{rnd}))/\sigma(W_{rnd}) \quad (3)$$

where W_S is the weight of found global alignment, W_{rnds} are weight of alignments along random sequences, M and σ are there mean value and dispersion.

The algorithm was applied in a program complex written in C language with computational cluster support (MPI). Obtained results were compared against genes annotations from GenBank.

Results and Discussion

Primary results were calculated by comparing obtained data on tRNA-like sequences layout against existing GenBank annotations (see Table 1).

Table 1. Number of tRNA-like sequences in various functional structures.*

	Bacterial	Invertebrates	Mammals	Patented	Phages	Plants	Primates	Rodents	Virulent	Vertebrates
Total sequence length, Mbp	***234,7***	***358,9***	***24,4***	***73,1***	***4,6***	***348,3***	***1 248,2***	***105,3***	***96,0***	***44,4***
Not described	478	1 130	174	872	26	909	282 097	909	54	239
repeat_region	8	65	157	0	0	47	211 829	1 243	1	79
gene	433	434	156	4	8	1 034	45 005	467	245	102
misc_feature	334	76	45	2	12	483	22 808	143	23	75
CDS	427	103	89	25	15	382	11 059	118	384	47
intron	7	27	117	1	0	149	5 482	318	10	18
exon	10	33	10	2	0	362	3 457	41	2	18
mRNA	14	19	65	0	1	32	1 702	31	18	3
prim_transcript	6	3	10	1	0	9	1 551	51	10	6
repeat_unit	0	20	89	0	0	15	756	51	0	131
mat_peptide	12	5	4	3	0	12	710	6	30	1
promoter	0	0	4	2	1	7	605	63	0	1
3'UTR	0	2	6	0	0	1	446	28	0	3
STS	0	0	0	0	0	0	400	0	0	0
rRNA	230	69	6	0	0	264	86	2	0	21
V_region	0	0	0	0	0	0	363	9	0	0
unsure	0	0	0	0	0	0	354	3	0	0
misc_RNA	174	2	0	0	0	11	15	0	24	5
5'UTR	0	0	2	0	0	4	138	6	0	24
D-loop	0	1	11	0	0	0	13	0	0	117
precursor_RNA	47	1	3	0	0	2	96	8	1	4
LTR	0	0	0	0	0	9	36	0	11	0
misc_signal	1	0	0	0	0	0	38	0	0	1
protein_bind	0	0	0	0	0	0	36	10	0	0
sig_peptide	0	0	0	1	0	1	37	0	0	0
satellite	0	0	11	0	0	2	23	2	0	0
scRNA	1	1	0	0	0	0	9	17	0	0
C_region	0	0	1	0	0	0	14	0	0	0
D_segment	0	0	0	0	0	0	10	0	0	0
misc_recomb	5	0	0	0	0	0	5	0	0	0
misc_structure	0	0	0	0	0	0	7	0	0	0
iDNA	0	1	0	0	0	1	6	1	0	0
primer_bind	0	1	2	0	0	5	2	0	0	0
enhancer	0	0	0	0	0	0	5	0	0	0
J_segment	0	0	0	0	0	0	5	0	0	0
snRNA	0	0	0	0	0	0	3	4	0	0
stem_loop	4	2	0	0	0	1	0	0	0	0
rep_origin	0	0	0	0	0	0	2	0	0	0

To calculate statistical significance of the obtained results series of experiments on random sequences was undertaken. Proceeding from these data a mean density of casual similarity cases was calculated. It was applied to every functional region described in GeneBank. Findings are shown in Table 2.

Table 2. Statistical significance of the results obtained (number of expected casual tRNK-like sequences among the found ones in the given functional area, %). *

	Bacterial	Invertebrates	Mammals	Patented	Phages	Plants	Primates	Rodents	Virulent	Vertebrates
mat_peptide	127,4	303,3	295,7	41,4	-	145,4	**3,2**	361,9	203,2	1177,1
mRNA	**37,6**	909,3	**18,6**	-	169,8	229,3	**4,5**	32,0	86,5	655,7
3'UTR	-	347,7	76,9	-	-	615,3	**3,6**	**14,3**	-	121,4
CDS	136,6	348,9	78,0	56,4	125,4	130,6	**2,6**	137,1	105,5	258,1
misc_RNA	**0,3**	393,4	-	-	-	133,7	**13,8**	-	**4,0**	**7,8**
sig_peptide	-	-	-	24,3	-	364,9	**10,7**	-	-	-
exon	-	**22,7**	135,0	183,2	-	**10,7**	**2,2**	16,6	249,5	37,8
rRNA	41,8	75,3	153,4	-	-	36,4	**5,2**	206,7	-	117,7
intron	-	**20,0**	**3,7**	220,5	-	28,4	**1,3**	**2,8**	67,4	40,7
gene	86,6	45,2	**11,4**	34,4	179,9	27,2	**1,2**	**6,5**	105,4	**16,8**
precursor_RNA	**0,7**	56,3	**8,1**	-	-	192,5	**3,8**	**10,1**	102,2	**7,1**
D-loop	-	151,7	109,1	-	-	-	89,1	-	-	**1,5**
C_region	-	-	193,9	-	-	-	**21,1**	-	-	-
repeat_region	**18,5**	**5,9**	**2,3**	-	-	62,2	**0,6**	**0,9**	134,0	**1,5**
primer_bind	-	118,5	25,3	-	-	**8,8**	70,0	-	-	-
iDNA	-	83,5	-	-	-	**6,7**	**19,6**	95,9	-	-
satellite	-	-	**14,1**	-	-	121,7	**12,6**	36,7	-	-
5'UTR	-	-	118,1	-	-	**21,7**	**5,1**	45,3	-	**1,7**
prim_transcript	38,1	66,4	**15,6**	**20,9**	-	40,7	**1,8**	**6,1**	69,6	41,4
promoter	-	-	44,4	59,1	59,2	**16,3**	**2,3**	**3,8**	-	68,9
LTR	-	-	-	-	-	63,1	**10,3**	-	99,4	-
misc_signal	104,3	-	-	-	-	-	**6,1**	-	-	33,7
misc_feature	**1,3**	**17,2**	**11,6**	93,9	**6,1**	**15,4**	**1,5**	**2,7**	**8,5**	**12,1**
V_region	-	-	-	-	-	-	**6,2**	94,5	-	-
stem_loop	**8,6**	**3,3**	-	-	-	88,2	-	-	-	-
ScRNA	61,7	27,9	-	-	-	-	**4,5**	**0,8**	-	-
repeat_unit	-	**7,0**	**1,0**	-	-	28,1	**1,0**	**1,1**	-	**0,2**
SnRNA	-	-	-	-	-	-	26,0	**4,5**	-	-
misc_recomb	**5,3**	-	-	-	-	-	**18,6**	-	-	-
protein_bind	-	-	-	-	-	-	**5,7**	**16,7**	-	-
D_segment	-	-	-	-	-	-	**13,0**	-	-	-
misc_structure	-	-	-	-	-	-	**8,6**	-	-	-
CAAT_signal	-	-	-	-	-	-	-	**5,3**	-	-
STS	-	-	-	-	-	-	**3,3**	-	-	-

* Data on known (marked in GenBank) tRNA regions are not shown to exclude its influence to the statistics. Almost all tRNA sequences tagged in GenBank was found by the algorithm used (unfound tRNA sequences appears too short to be detected by the program).

The method applied allows to detect very diverged tRNA successors. It reveals weaker sequence similarity than standard BLAST and PSI-BLAST algorithm that was shown on MIRs repeat research (Korotkov, 2000). Existing works on revealing tRNA-like sequences were based on tRNA-specific canonical sites (Marvel, 1986) were not able to detect other evolutionary altered tRNAs.

Further detailed analysis of the results is being taken nowadays.

Acknowledgements

This work was supported by the grant "Designing computer programs for analysis of structural and functional genome properties" of Russian Ministry of Industry, Science and Technologies.

References

1. Korotkov E.V., Korotkova M.A. (2000) MIRs: family repeats that is common for many vertebrates. Mol. Biol. (Russian). 34, 348-353.
2. Mans R.M., Pleij C.W. etc. (1991) tRNA-like structures. Structure, function and evolutionary significance. Eur. J. Biochem. 201(2), 303-24.
3. Marvel C.C. (1986) A program for the identification of tRNA-like structures in DNA sequence data. NAR. 14(1), 431-5.
4. Sibbald P.R., Argos P. (1990) Weighting aligned protein or nucleic acid sequences to correct for unequal representation. J. Mol. Biol. 216(4), 813-8.
5. Waterman M.S. Introduction to Computational Biology. Map Sequences and Genomes. Chapman and Hall Press, London, 1995.

A GENETIC ALGORITHM FOR THE INVERSE FOLDING PROBLEM OF RNA

[1] Titov I.I., [2] Pal'yanov A.Yu.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: titov@bionet.nsc.ru
[2] Novosibirsk State University, Novosibirsk, 630090, Russia
*Corresponding author

Key words: *secondary structure, RNA, artificial evolution, genetic algorithm, nanotechnology*

Resume

Motivation: Solution of the inverse folding problem of RNA is important for understanding the laws of spatial RNA structure coding. A search for RNA sequences that form a specific structure is of practical importance for nanotechnology, namely, for the production of nanoscale devices.

Results: A genetic algorithm for reconstructing RNA sequences from a specific secondary structure was developed. The algorithm was tested on calculations of RNA sequences with (a) stable and (b) unstable secondary structures. The test results showed that the algorithm yielded the best sequence from approximately 10^{26} sequences. We calculated the RNA structures that were able to form artificial structures, namely (c) a cube and (d) a square lattice. The size of the structures (c) and (d) is smaller than that of DNA analogues synthesized earlier (Seeman, 1985; Chen, Seeman, 1991) and further decrease in size is unlikely.

"The principles of physics…do not speak against the possibility of maneuvering things atom by atom. It is not an attempt to violate any laws; it is something, in principle, that can be done; but in practice, it has not been done because we are too big." *R.P. Feynman* There's Plenty of Room at the Bottom "Every living thing is made of cells that are chock full of nanomachines." R.E. Smalley

Introduction

Because of a specific nature of complementary interactions and the modern manipulation techniques, nucleic acids are amongst the most promising materials for design of nanoscale devices with predictable properties. The production of such devices employs a many million-year experience of the evolution of living beings: it is possible to synthesize a DNA lattice whose elementary cell is analogous to DNA-junction of living cell (Seeman, 1985). However, the evolution could hardly exhaust all the possible sequences; therefore, it would be sufficient to follow the common principles of molecular biology, and the complex DNA structures could be created (Seeman, 1985; Chen, Seeman, 1991).

The artificial RNA/DNA structures have a large variety of practical applications: substrates for molecular electronics, templates, sensors, springs, etc. A very important characteristic of RNA is its capability to fold spontaneously into unique structure. Although the question of how this ability is coded in a symbolic sequence has been extensively studied for several decades, it still remains vague. Therefore, the reconstruction of a sequence from a spatial RNA structure (i.e., the inverse problem of folding) is a basic problem of great practical importance. One way to solve this problem is to refer to the natural analogues and conduct an evolution in a tube following the so-called SELEX protocol. This protocol implies selection of the most functionally active artificial RNAs or peptides, for example, those with high affinity for ligands. A similar technology for computer modeling called a genetic algorithm has been successfully used in the optimization of complex systems. This paper describes the genetic algorithm for the solution of the inverse folding problem.

Methods

The thermodynamic parameters of the secondary structure were taken from (Turner, Sugimoto, 1988).

The secondary RNA structure was calculated by our algorithm *GArna* (Titov et al., 2002; Vorobiev et al., 2002).

Nonrandomness of the sequences obtained was assessed by calculating a relative stability of the secondary structure (Z score; Titov et al., 2000).

The optimization quality was assessed on test samples by calculating the degree of deviation of the structural energy from a linear dependence on sequence length (Titov et al., 2000).

Algorithm

The genetic algorithm protocol involved the following steps:

1. Generation of the initial population from sequences of random composition. Then, the population evolved under the cyclic action of genetic operators: selection, recombinations and mutations.

2. Stochastic selection of individuals from the population according to their fitness values.

3. Filling of the vacancies formed after step 2 by recombination results.

4. A part of the population was subjected to point mutations, which were implemented as conventional algorithms of a stochastic search (Metropolis et al., 1957).

5. Calculation of the dispersion of individual fitness values in the population. The calculations were aborted if the value of the dispersion was smaller than the corresponding threshold value, which evidenced the convergence of the algorithm; otherwise, we repeated from Step 2.

Results and Discussion

To estimate the optimization quality, the algorithm was tested on the following simple examples.

(a) Selection towards a stable secondary structure

A hairpin consisting of a loop of an optimal size and a stem of G and C tracts represents the most stable structure. The energy of such a structure increases linearly with a stem length, which can be used as a criterion of optimization quality. In the calculations, the linear dependence held until the sequence reached 54 nucleotides in length. A typical evolutionary process involved an increase in the GC content and ordering of the sequence. For a length of 54 nucleotides, the Z-score was significant and equaled –13. The sequence calculated was close to optimal. Let us estimate an optimization ability of the algorithm as a ratio of the number of all the possible sequences of a length of 54 nucleotides ($4^{54} = 3.2 \times 10^{32}$ sequences) to the number of the sequences more stable that the sequence found (the latter had fewer G and C tracts; totally, there were 3.7×10^6 such sequences). Thus, the optimization ability of the algorithm was 10^{26}.

(b) Selection towards an unstable secondary structure

In this case, the resulting sequences should be those lacking complementary partners, such as polyA and others. The proposed algorithm allowed quick solutions of this problem. Then, we restricted our search by the sequences of G and C nucleotides with equal contents so that the result was less obvious. Due to this restriction, a significant part of the sequence space became unavailable, which reduced the convergence. As a result, the evolutionary process proceeded exclusively via nucleotide ordering, and the resulting value of the Z score was significant and equal to 4.6. The final sequence consisted of CCG-triplets and G-tracts.

In comparison with the case (a), the simplicity of optimization for the case (b) agrees with our previous result, which shows that sequences with a stable secondary structure are rare in the space of all the possible sequences (Titov et al., 2002). Interestingly, in the context of information theory, the sequences obtained from tests (a) and (b) are very simple because they encode "simple messages" about the secondary structure: "the most stable structure" (a) and "the most unstable structure" (b).

The following examples of application of the algorithm belong to the field of biotechnology.

(c) Calculation of the RNA sequences forming a square lattice

Square DNA lattices were obtained earlier at Seeman's laboratory (Seeman, 1985). Let us briefly describe the experimental procedure. At first, short DNA chains were synthesized, four of these chains forming a 4-arm branched junction—an elementary lattice cell. The nucleotide chain composition was selected using a PC to avoid alternative pairing. The arms had "sticky ends", which bound the junctions into a lattice by complementary interactions. The subsequent linking by ligases created covalent links between the elementary cells.

Accurate selection of a nucleotide sequence of the junction is important for increase in the yield of the product. A straightforward stabilization of the target structure by increasing the GC content might cause formation of alternative structures with the stabilities similar to that of the target structure due to a small size of the alphabet. Thus, the most appropriate way is to maximize the thermodynamic probability of the implementation of the target structure, i.e. to increase its thermodynamic gap ΔE (Fig. 1). Using ΔE, we selected samples for the cases (c) and (d) of this paper. Finally, the gap was significant for the junction ($\Delta E = 16.3$ kcal\mole), and the lattice period was halved compared to that of the DNA lattice (Fig. 2a). Figure 2 shows a junction with half-turned double strands; obviously, it is impossible to obtain a lattice with a shorter period.

(d) RNA cube design

Previously, a DNA cube was obtained in a rather complex manner from several chains (Chen, Seeman, 1991). It would be interesting to obtain a cube formed by spontaneous folding of a single RNA strand. This is possible if pseudoknot contacts are present in the secondary-structure trefoil (Fig. 3). A search for an RNA cube sequence was similar to example (c), when we maximized the width of the energy gap ΔE. Since the thermodynamic table is not available for pseudoknots, we used a table of secondary structure interactions. Despite this obvious simplification, the thermodynamic gap obtained was so wide (30 kcal\mole) that the instability of pseudoknot contacts compared with the common secondary-structure energy could not have exceeded the gap width. Nevertheless, we plan to employ more consistent approach and test our results with molecular force field calculations and take into account sterical restrictions. The RNA cube can be used as a matrix assembly of protein complexes (placed on the cube sides), a ligand cage, an inactive ribozyme state, etc.

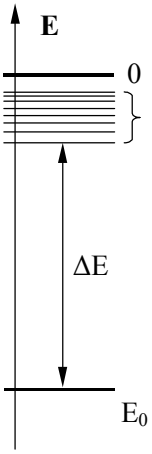


Fig. 1. The range of RNA secondary structure states includes a global minimum E_0 and a great number of alternative states. A wide thermodynamic gap ΔE guaranties a high probability for the equilibrium molecule to have an optimal conformation.

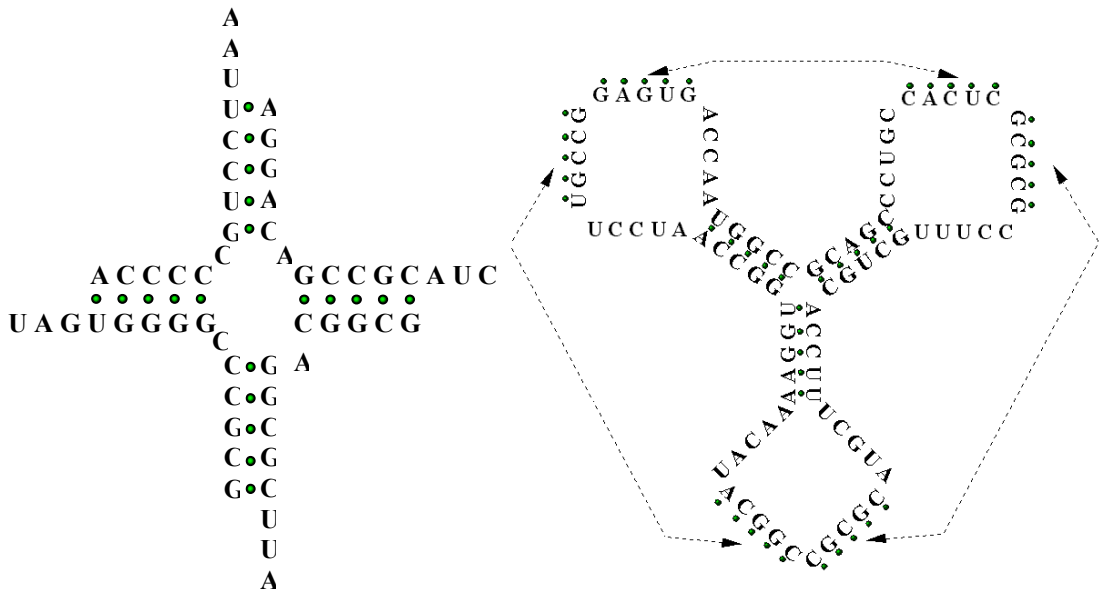


Fig. 2. The calculated sequences of RNA junction (left) and RNA cube (right). The cube is formed by pseudoknot contacts of loops (arrows).

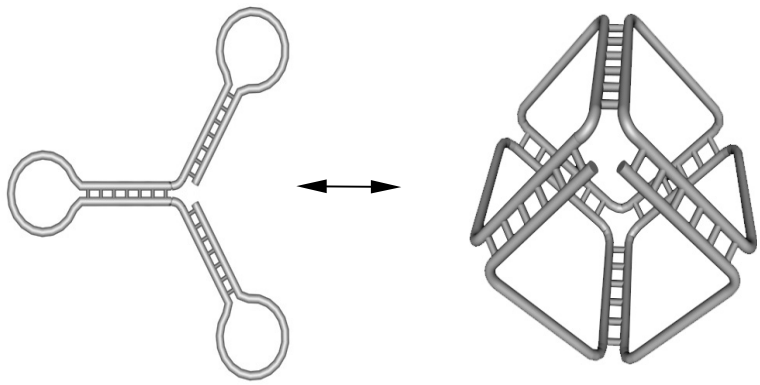


Fig. 3. Model of the formation of a nanocube from an RNA: first, the secondary trefoil structure is formed (left), which through three pseudoknot contacts forms an isomorphic cube in the space (right, artist view). The cube has 6 single-stranded and 6 double-stranded edges. See also Fig. 2.

Acknowledgements

The work was supported by the Siberian Branch of the Russian Academy of Sciences (Integration Project № 65) and INTAS (grant № 2001-2126).

References

1. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.N., Teller E. (1953). J. Chem. Phys. 21:1087.
2. Seeman N.C. (1985). Macromolecular design, nucleic acid junctions and crystal formation. J. Biomol. Struct. Dynam. 3(1):11–34.
3. Chen J., Seeman N.C. (1991). Synthesis from DNA of a molecule with the connectivity of a cube. Nature. 350: 631-3.
4. Turner D.H., Sugimoto N. (1988). RNA structure prediction. Ann. Rev. Biophys. Chem. 17:167–192.
5. Titov I.I., Ivanisenko V.A., Kolchanov N.A. (2000). FITNESS—A WWW-resource for RNA folding simulation based on genetic algorithm with local optimization. Comput. Technol. 5:48–56.

A DATABASE ON ALTERNATIVE SPLICE FORMS ON THE INTEGRATED GENETIC MAP SERVICE (IGMS)

**** Pospisil H., Herrmann A., Pankow H., Reich J.***

Max-Delbrueck-Center for Molecular Medicine, 13125 Berlin, Germany, e-mail: pospisil@mdc-berlin.de

*Corresponding author

Key words: *alternative splicing, ESTs, database, gene expression profiles, colon cancer*

Resume

Motivation: Many databases are available on the Internet regarding DNA sequences, which focuses on different biological or medical properties. This situation makes the search for specific (often linked) information difficult with regard to a) the completeness and b) the specificity of the received information. The IGMS is a comprehensive information system that combines the knowledge from genomic sequence, genetic map and genetic disorders databases. This system is updated weekly and focuses on the analysis of EST data.

Results: The IGMS identifies UniGene clusters that are differentially expressed in different types of cancer with respect different reference tissues. The results can be combined with clinical data to asses the potential relevance of specific genes for patient survival or metastatic spread. The second application maps EST with a specific expression profile. Our third application generates a database of alternative splice forms for eight organisms from EST and mRNA sequence data. The results can be used to find splicing patterns specific for certain tissues or tumour types.

Availability: http://medseq.bioinf.mdc-berlin.de/imap/ or http://www.bioinf.mdc-berlin.de/

Introduction

There are over 15 millions sequence records at the GenBank (May, 2002 [1]), including more than 11 million expressed sequence tags (ESTs) (May, 2002, [1]), representing a wide variety of different organisms, tissue types, including diseased and normal cell lines. This large amount of data raises a lot of possible questions for investigations the complexity of being.

We present an online available, weekly updated service that combines as well the knowledge from genomic sequence, genetic map and genetic disorders databases as a database of potentially alternatively spliced forms. The *Integrated Genetic Map Service* (IGMS) system focuses on the analysis of EST data and enables to extract sequences of interest from the large amount of entries.

It is available under http://medseq.bioinf.mdc-berlin.de/imap.

Resources

The IGMS system has integrated the following database resources:

GeneMap99 (GeneMap99_gb4 / GeneMap99_sg3). This human gene map is the result of the collaboration of an International RH Mapping Consortium. Two Radiation Hybrid (RH) panels were used: the Genebridge4 (GB4) panel and the Stanford G3 panel. GB4 provides long-range map continuity while G3 gives higher local resolution. The GeneMap'99 represents the location of more than 30,000 genes. It's accessible from NCBI.

WHI. This is the final STS-based genetic linkage map resulting of the Human Physical Mapping Project at Whitehead Institute/MIT Genome Center.

MGD. The Mouse Genome Database at Jackson Laboratory.

HMGD. The Human-Mouse Genome Database at Jackson Laboratory.

LDB. The Genetic Location Database (LDB) gives locations for expressed sequences and polymorphic markers. Locations are obtained by integrating data of different types (genetic linkage maps, radiation hybrid maps, physical maps, cytogenetic data and mouse homology) and constructing a single 'summary' map.

OMIM. The Online Mendelian Inheritance in Man. This database is a catalog of human genes and genetic disorders.

GenBank. The GenBank sequence database at NCBI.

RefSeq. The NCBI reference sequences (RefSeq) provide standards for complete genomic nucleic acids, assembled contigs, transcripts and proteins. RefSeq records are derived from GenBank and the literature to provide a non-redundant set of sequences that facilitate sequence identification and information retrieval. (Note: The IGMS system has included the human XM_* (e.g. XM_066987) mRNA sequences only. These sequence records represent genes of unknown function).

UniGene. The UniGene database is a collection of non-redundant sequence clusters derived from known genes, ESTs and their high scored GenBank sequence homologies. It must be noted, that the UniGene project did not attempted to build an overlapping consensus sequence or contig.

CGAP. All EST libraries for human and mouse created by the Cancer Genome Anatomy Project at NCBI was copied into the IGMS system.

Affymetrix GeneChips. The GeneChips HU95A, HU95B, HU95C, HU95D, HU95E of Affymetrix Inc.

Additionally we integrated an Alternative Splice Database. This database represents splice forms of Arabidopsis thaliana, Bos taurus, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus, Rattus norvegicus and Homo sapiens from ESTs and mRNA GenBank sequence records. Our algorithm [2-4] defines a possible alternative splice form by comparing high-scoring ESTs to mRNA sequences using BLAST. Filtering programs compare the ends of each aligned sequence pair for deletions or insertions in the EST sequence, which suggest the existence of alternative splice forms.

Implementation

The IGMS is divided into four major parts:

GenBank part

This function copies complete information from the GenBank [5] source. Optionally you can extract the complete GenBank entries, only the references or only the CDS or exon feature sequences. The latter function enables one to create a own FASTA formatted sequence database.

Gene Expression part

Compared with other websites, you can find some special functions in the information class "Gene Expression Profiles", which are very useful for the analysis of gene expression, as e.g. tissue histology and Affymetrix GeneChip probe set numbers. For that case we have integrated the complete UniGene cluster sets collection and the CGAP EST libraries created by the Cancer Genome Anatomy Project at NCBI into the IGMS system.

As a new function it is now possible to identify UniGene clusters that are differentially expressed in different types of cancer with respect different reference tissues, using for example, as criteria defined ratios of the number of ESTs found in tumour tissues as compared to the number found in normal tissues and a defined number of ESTs per cluster. It could be very interesting to retrieve e.g. *all human UniGene clusters with at least 30 ESTs, more than 90% are found in cancerious tissues and at least one EST must expressed in a specific tissue (e.g. in colon)*. The results can be combined with clinical data to asses the potential relevance of specific genes for patient survival or metastatic spread.

Integrated Genetic maps

This part of the IGMS system maps EST with a specific expression profile, e.g. representing genes over expressed in breast cancer, to the corresponding regions of the genome and vice e versa, e.g. maps all genes on chromosome 8 that are over expressed in breast cancer.

Alternative Splice forms

In the alternative splice database a list of all possible alternative splice forms for seven organisms is available. This list consists the direct link to the information mentioned before. Furthermore it is possible to search for a specific mRNA sequence to check if this sequence could be alternatively spliced.

Furthermore we added a function to summarize all available information for a selected sequence (see Fig. 2).

Results

The IGMS was used to investigate alternative splicing in colon cancer. For that we examined all human UniGene clusters with at least 30 ESTs per cluster by using the "*gene expression profile*" function. Afterwards all sequences expressed in colon were selected from these 13.954 different clusters ("*extract ESTs+Tissue+Chr*"). As second step we used our human alternative splice database that was created as described in [2] (the average alignment identity is at least 98% at it each HSP is at least 30 bp long). We found that 1707 colon ESTs indicate alternative splicing of 2857 different mRNAs.

One example is shown in Fig. 1. (The complete table is available at http://www.bioinf.mdc-berlin.de/splice/colon/).

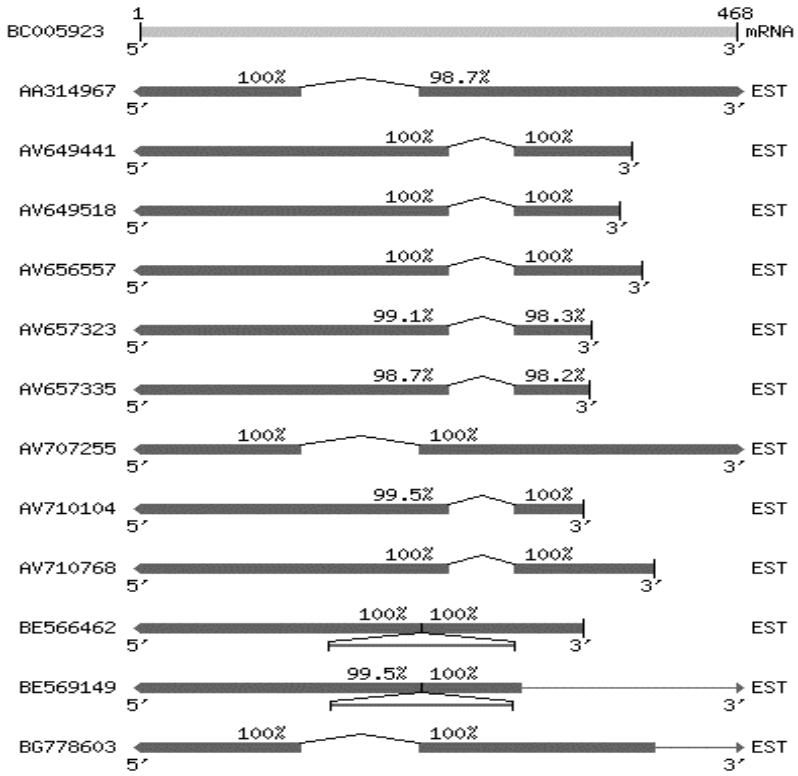
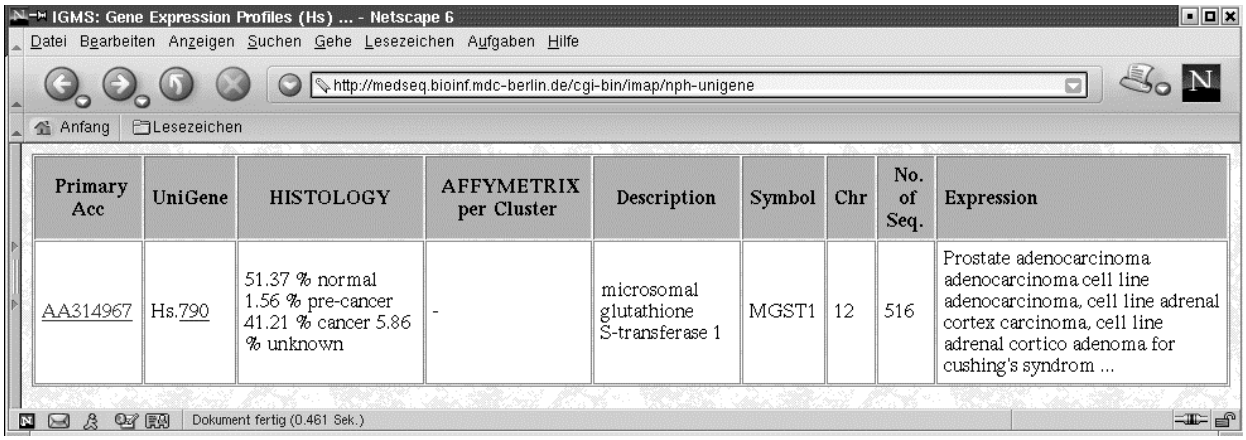


Fig. 1. All possible alternative splice forms of mRNA BC005923 (microsomal glutathione S-transferase 1). mRNA is indicated in light grey, the ESTs in dark grey. The alternative splice form is indicated by a deletion of >30bp (e.g. AA314967) or an insert of >30bp (e.g. BE566462). The EST AA314967 is derived from a HCC cell line in colon and belongs to UniGene cluster Hs.790 (see Fig. 2).



Primary Acc	UniGene	HISTOLOGY	AFFYMETRIX per Cluster	Description	Symbol	Chr	No. of Seq.	Expression
AA314967	Hs.790	51.37 % normal 1.56 % pre-cancer 41.21 % cancer 5.86 % unknown	-	microsomal glutathione S-transferase 1	MGST1	12	516	Prostate adenocarcinoma adenocarcinoma cell line adenocarcinoma, cell line adrenal cortex carcinoma, cell line adrenal cortico adenoma for cushing's syndrom ...

Fig. 2. Overview of the UniGene cluster Hs.790. The IGMS summarizes the histology information, gives the Affymetrix GeneChip probe set number (if available), the description of the cluster, genetic symbol, location of the chromosome, the size of the cluster and the corresponding tissue types.

Discussion

We present here a retrieval system to extract genomic data in combination with knowledge from genomic sequences, genetic map and genetic disorders databases and a database of potentially alternatively spliced forms. This approach enables one to filter out the sequences and information of interest very selectively.

The IGMS includes three novel functions: (a) identification of UniGene clusters that are differentially expressed, (b) mapping ESTs with a specific expression profile and (c) a database of alternative splice forms for seven organisms.

The main advantage of this system is the combination of genomic information with expression information and with alternative splice information.

References

1. GenBank and dbEST Overview 2002 http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html
2. Brett D., Kemmner W. et al. (2001) A rapid bioinformatic method identifies novel genes with direct clinical relevance to colon cancer. Oncogene. 20, 4581-5.
3. Brett D., Hanke J. et al. (2000) EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Lett. 474, 83-6.
4. Brett D, Pospisil H. et al. (2002) Alternative splicing and genome complexity. Nat Genet. 30, 29-30.
5. Benson D.A., Karsch-Mizrachi I. et al. (2002) GenBank. Nucl. Acids Res. 30, 17-20.

ARE PATTERNS OF ALTERNATIVE SPLICING OF MAMMALIAN GENES CONSERVED?

[1]* Nurtdinov R.N., [2] Artamonova I.I., [3] Mironov A.A., [3] Gelfand M.S.

[1] Moscow State University, College of Physics, Department of Biophysics, GSP-2, 119922, Moscow, Russia,
e-mail: n_ramil@mail.ru
[2] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry RAS, 117871, Moscow, Russia
[3] State Scientific Center GosNIIGenetika, 113545, Moscow, Russia
* Corresponding author

Key words: *computer analysis, alternative splicing, exon-intron structure*

Resume

Motivation: Alternative splicing is a major mechanism of generating protein diversity in mammalian genomes. Up to 60% human genes have alternative variants of mRNA. To understand the evolution of alternative splicing, it necessary to compare variants of alternative splicing in related genomes.

Results: We analyzed conservation of alternative splicing patterns in orthologous genes from the human and mouse genomes. Our results demonstrate considerable diversity of alternative splicing in these genomes. Orthologous genes with different alternative splicing patterns are good candidates for the role in species-specific development.

Introduction

One of the essential features of eukaryotic genes is that one gene can produce many isoforms of mRNA. The significance of alternative splicing was fully appreciated after completion of the draft human genome. It was found that the genome contains 30-35 thousands genes (International Human Genome Sequencing Consortium, 2001), unlike the earlier estimates that have been as high as 120 thousand genes (Liang et al., 2000). The proportion of alternatively spliced genes (we allow ourselves some liberty of speech, applying the term "splicing" to genes, although in reality this process operates on pre-mRNA transcripts) was estimated to reach 35-60% (Venter et al, 2001). Thus it is of interest to compare the alternatively spliced variants of orthologous mammalians genes, in our case, from mouse and human.

Methods

Alternatively spliced mouse genes from AsMamDB (Zhou et al., 2001) were used to scan the draft human genome (International Human Genome Sequencing Consortium, 2001) using BLAST. The mouse protein isoforms were aligned to orthologous genes using Pro-Frame (Mironov et al., 2001), and thus the exon-intron structure of the human genes was established. Proteins isoforms that could not be aligned throughout their length were analyzed in detail using the EST alignments from the Human Alternative Splicing Database (HASDB) (Modrek et al., 2001) and the published experimental data from the literature.

Results

We compared 62 pair of alternatively spliced human and mouse ortholog pairs. 79 "primitive" alternatives in mouse genes were as follows: 37 cassette (on/off) exons, 6 retained introns, 3 pairs of alternative (mutually exclusive) exons, 16 alternative donor sites and 17 alternative acceptor sites. Ten alternatively spliced mouse genes had variants not observed in the human genome. The most interesting cases of non-conserved alternative splicing are presented below. Common variant(s) are placed in between the genome lines under "Coinciding variant(s)". Genome-specific variants are marked "Mouse-specific Variant(s)" and "Human-specific Variant(s)". Dotted lines indicate hypothetical prolongation of mRNA (EST) when it has broken after stop-codon.

The mouse interleukin 4 receptor gene has three alternative mRNAs (Fig. 1). Cassette exon 5a and retained intron contain in-frame stop-codons. The human gene also has cassette exon 5a, but it is not homologous to the mouse one. There is no region homologous to the retained intron. In addition, HASDB has a variant with cassette exon 3a containing an in-frame stop-codon and an alternative acceptor site in human exon 5. Both these alternatives have no mouse counterparts.

The mouse glucocorticoid-induced TNFR family-related protein gene has four mRNAs (Fig. 2). Both human and mouse genes have cassette exon and an alternative acceptor site. On the other hand, the human analog of the mouse retained intron is not homologous to the mouse one and their lengths differ by a non-integer number of codons. HASDB analysis produced an alternative donor site in human cassette exon 3a, not conserved in the mouse gene.

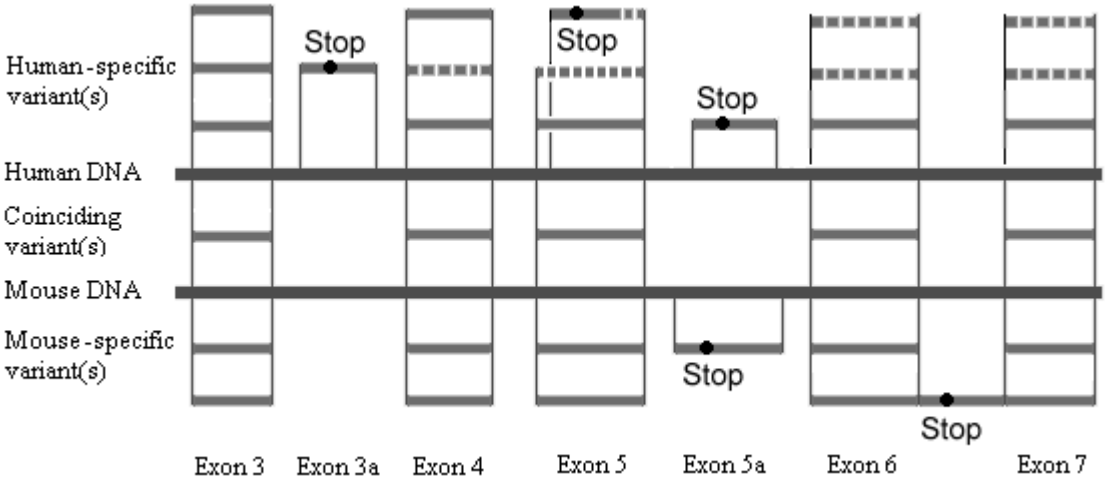


Fig. 1. Splicing of the interleukin 4 receptor gene.

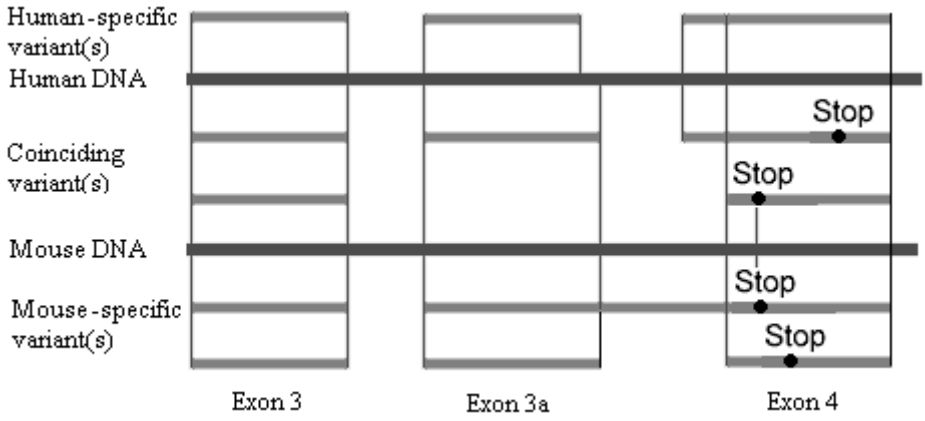


Fig. 2. Splicing of the glucocorticoid-induced TNFR family-related protein gene.

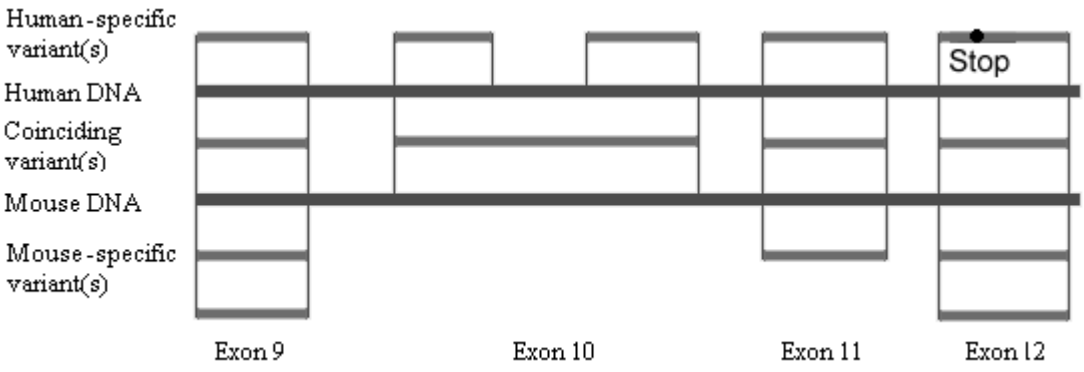


Fig. 3. Splicing of the autoimmune regulator gene.

The autoimmune regulator gene of mouse has four alternatives, two cassette exons and alternative donor and acceptor sites. The human gene has both cassette exons and the acceptor site, but not the donor site. HASDB also has a variant with a retained intron within the cassette exon (Fig. 3).

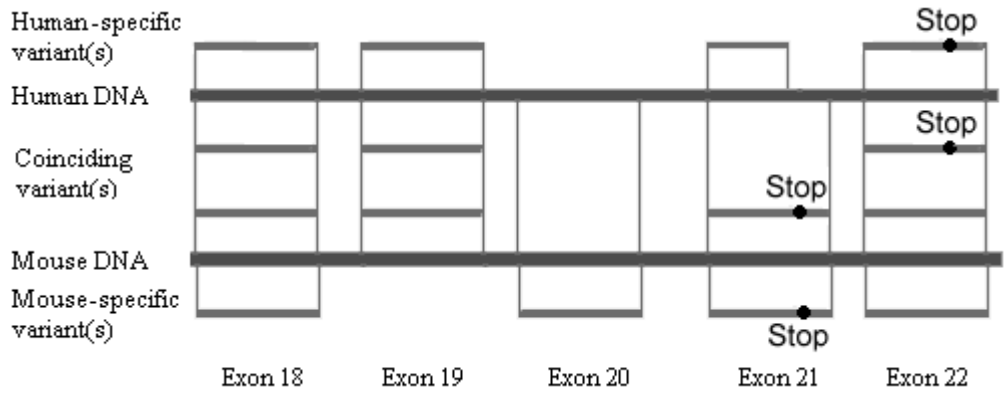


Fig. 4. Splicing of the smoothelin gene.

The smoothelin gene of mouse has three cassette exons (Fig. 4). All three exons were found in the human gene. However, HASDB contains a variant with an alternative donor site in exon 21 that is not found in the mouse gene. This donor site is found before terminated codon that allows to translate part of next exon.

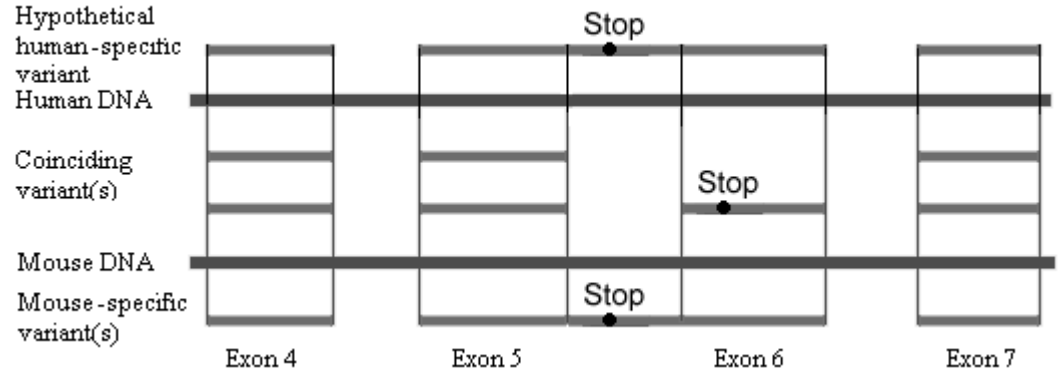


Fig. 5. Splicing of the FMS-like tyrosine kinase 3 ligand gene.

The mouse FMS-like tyrosine kinase 3 ligand gene has three alternative mRNAs (Fig. 5). The human gene has the conserved cassette exon with an in-frame stop-codon. The human analog of the retained intron is not homologous to the mouse one and it is twice longer that the latter. But there is a stop-codon in the human gene at approximately the same position as in the mouse gene, and thus the protein length is the same in both cases.

Discussion

Although preliminary, these results demonstrate a considerable diversity of alternative splicing in the human and mouse genomes. Orthologous genes with different alternative splicing patterns are good candidates for the role in species-specific development and speciation in general. As genomic comparison supplemented by analysis of EST seems to be a powerful tool, this pilot study will be continued using larger-scale genome-EST analysis.

Acknowledgements

This work was supported by grants from RFBR (00-15-99362), INTAS (99-1476), HHMI (55000309) and LICR (CRDF RB0-1268). We are grateful to A.Baranova and V.Makeev for useful discussion.

References

1. Ji H. Zhou, Q. Wen, F. Xia, H. Lu, X. Li. (2001) AsMamDB: an alternative splice database of mammals. Nucl. Acids Res. 29, 260-263.
2. International Human Genome Sequencing Consortium. (2001) Initial sequencing and analysis of the human genome. Nature. 409, 860-922.

3. Liang F., Holt I., Pertea G., Karamycheva S., Salzberg S. L., Ouackenbush J. (2000) Gene Index analysis of the human genome estimates approximately 120.000 genes. Nat Genet. 25, 239-240.
4. Mironov A.A., Novichkov P.S., Gelfand M.S. (2001) Pro-Frame: similarity-based gene recognition in eucariotic DNA sequences with errors. Bioinformatics. 17, 13-15.
5. Modrek B., Resch A., Grasso C., Lee C. (2001) Genome-Wide Detection of Alternative Splicing in Expressed Sequences of Human Genes. Nucl. Acids Res. 29, 2850–2859.
6. Venter J.C. et al. (2001) The sequence of the human genome. Science. 291, 1304-1351.

DATABASE ON mRNA-LOCATED EUKARYOTIC TRANSLATIONAL SIGNALS

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

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Tsukuba Institute, Institute of Physical Chemical Biology (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074 JAPAN
*Corresponding author: ak@bionet.nsc.ru

***Key words**: translational signals, database, mRNA, enhancer*

Resume

Motivation: It is known that signal sequences found in eukaryotic mRNAs control their translation rate, cytoplasmic stability, or intracellular localization. Development of computational tools for finding and analyzing them will be important both for basic research and application to genetic engineering.

Results: We developed the database on translational signals (TRSIG), collecting information on mRNA fragments with experimentally observed specific activities, including enhancers (general, tissue-, stage-, and stress-specific), determinants of mRNA cytoplasmic stability, internal ribosome entry sites (IRES), etc. The database consists of four parts (signal descriptions, signal sequences, experiments, and experimental full-sized sequences) combined on the SRS platform. It is also supplied with a BLAST search to detect local homologies between annotated signals and user-defined mRNA.

Availability: The database is available at *http://wwwmgs.bionet.nsc.ru/mgs/dbases/trsig/* as a part of the computer system mRNA-FAST (Kochetov et al., 2001). It is also available at http:/www.rtc.riken.go.jp/jouhou/trsig/trsig.html.

Introduction

It is well known that control of eukaryotic gene expression may occur at post-transcriptional level and expression signals are often located in mRNA sequences (within 5'UTR or 3'UTR). Most translational signals presently known have not been characterized in detail: commonly, they were described as mRNA fragments with some specific activities. It is likely that folding of RNA molecule into various conformations hampered the research of mRNA-located signals.

We have developed a database compiling the sequences of mRNA-located expression signals and published experimental data on their specific activities. The database (TRSIG) is installed on SRS platform and may be accessed through Internet. It is supplied with a BLAST program allowing the user to search for the local homology between annotated translational signals and the mRNA of interest. TRSIG also collects sequences of signal-containing full-sized experimental mRNAs together with the data on their relative activities. It provides a possibility to analyze signals in more detail, including their secondary structure.

Database Description

TRSIG is implemented on SRS platform (http://www.lionbio.co.uk) and consists of four databases (Fig.).

Database on objects (A). This database compiles the information on translational signals (fields LOCATION, TYPE, and COMMENT), their taxonomy (OC and OS, similar to an EMBL entry), and main characteristics, namely, the presence of cap (CAP) and poly(A) tail (POLYA). Each entry contains nucleotide sequence (field SQ) and references to the other TRSIG subdatabases (LINK). An important field EXPERIMENT contains identifiers of the database on experiments (B), where experiments on this signal are described.

Database on experiments (B) contains the description of experiments made with translational signals. It is linked to IDs of the corresponding object database entries (OBJID), references to the published data (REFERENCE), and experiment descriptions (TYPE, CELL, and COMMENT). The field ACTIVITY contains experimental data, including identifiers of entries in the sequence database (C) and their experimentally measured activities. These data may be used for comparative analysis of signal-containing sequences and detection of the essential characteristics of regulatory regions.

Database on nucleotide sequences of regulatory regions (C) compiles nucleotide sequences of mRNA regions with regulatory activities demonstrated in model experiments. It also contains cross-references to the corresponding entries of the databases on objects (A) and experiments (B), information on the availability of full-sized mRNA version (field LONG, also see D). The field EXPR contains experimentally measured activities of the sequences described; e.g., the activity of S0097 (relative to S0098-S0102) is 57 (Fig. C).

Database on full-sized mRNAs (D) contains full-sized sequences of mRNA molecules described in the original published data, including the regulatory regions annotated in the database (C). The complete experimental sequence allows the

secondary structure and long-range complementary interactions to be taken into account. Each entry contains also comments on the sequence and positions of the translation start (START) and stop (STOP) codons.

ID ADHZM5 **(A) OBJECT**
DATE 20010321
AUTHOR KOCHETOV
LOCATION 5'UTR
TYPE stress-specific enhancer
OC Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; euphyllophytes; Spermatophyta; Magnoliophyta; Liliopsida; Poales; Poaceae; Zea.
OS Zea mays
GENE ADH1, alcoholdehydrogenase I
CAP capped
POLYA polyadenylated
SQ attttctcgctcctcacaggctcatctcgtttggatcgattggtttcgtaactggtgaaggactgagggtctcgg…
COMMENTSEQ 5'UTR of ADH1 gene mRNA
KEYWORD enhancer, hypoxia, anoxia, anaerobiosis, stress
COMMENT It was found that translation of alcoholdehydrogenase mRNA was efficient …
LINK EMBL_AC X00580
EXPERIMENT E0066

ID E0066 **(B) EXPERIMENT**
OBJID ADHZM5 ADHZM3
REFERENCE Bailey-Serres J., Dawe R.K. Both 5' and 3' sequences of maize adh1 mRNA are required for enhanced translation under low-oxygen conditions. Plant Physiol. 1996. 112. 685-695.
TYPE transient expression of capped polyadenylated mRNAs in plant protoplasts
CELL maize
COMMENT Translational efficiencies of reporter mRNAs containing UTR sequences …
ACTIVITY S0101=0.1 S0097=57 S0098=37 S0099=20 S0100=17 S0102=16

ID S0097 **(C) SIGNAL SEQUENCE**
OBJID ADHZM5 ADHZM3
SQ atagggagaccgaattcgagctcattttctcgctcctcacaggctcatctcgtttggatcgattggtttcgtaac…
LONG YES
COMMENT Design of mRNA 5'UTR of GUS reporter gene: first 23 nt were taken from…
EXPR E0066=57

ID S0097 **(D) FULL-SIZE mRNA**
COMMENT design of reporter GUS gene mRNA 5'UTR in experiment: first 23 nt were..
START 145
STOP 2028
SQ atagggagaccgaattcgagctcattttctcgctcctcacaggctcatctcgtttggatcgattggtttcgtaac…

Fig. Examples of entries of TRSIG subdatabases: (**A**) object description; (**B**) experiment description; (**C**) description of sequence of regulatory region (**D**); and full-sized mRNA sequence.

Implementation

TRSIG can be used to find local homologies between a set of post-transcriptional signals and an mRNA of interest to hypothesize its expression control. The simplest way is to use BLAST module to search for homology between an untranslated region of interest and the sequences compiled in TRSIG (Fig. C). To test it, we tried to find a local homology (longer than a 10-nucleotide identity stretch) between the sample of 184 full-sized dicot plant 5'UTRs and 119 TRSIG-annotated sequences. It was found that some 5'UTRs contained motifs characteristic of translational enhancers.

For example, 5'UTRs of glycoprotein P gene of Arabidopsis thaliana (ATATPGP1) and S1 ribosomal protein of Spinacea oleracea (CLSORPS1G) contain fragments (**accaacaacaac** and **caacaaacaacaa**, respectively) of the strong translational enhancer omega (5'UTR of tobacco mosaic virus).

```
ATATPGP1 (145 nt): ..cataacaccaacaacaactcacgaagctccagagaaactcaccggaaATG
                              ||||||||||||
TMV:            ..tacaacaattaccaacaacaacaaacaacaacaacacattacaattactatttaca
                                        ||||||||||||
CLSORPS1G (91 nt):...cttatctgcgtatctcaacaaacaacaacataggaagaagatcaaagagtagc
```

It is known that this (caa)-enriched omega fragment is capable of increasing translation efficiency of reporter mRNAs in model experiments.

The 5'UTR of lti78 arabidopsis gene (ATLTI78), which expression is induced by low temperature, contained a fragment of the translational enhancer of tobacco etch virus (**tacttctattg**).

```
TEV:                ..cattctacttctattgcagcaatttaaatcatttcttttaaagcaaaagcaattt
                                  ||||||||||||
ATLTI78 (81 nt): ..tttgatacttctattggaaagaaaaaaatctttggaaaATG
```

Notably, this fragment represents an active domain of TEV enhancer and could render experimental mRNAs able to initiate translation in a cap-independent manner (Niepel, Gallie, 1999). It may be assumed that cap-independent translation increases the translation efficiency of this mRNA under unfavorable cold shock conditions. We believe that analysis of eukaryotic mRNAs for the presence of enhancer-like fragments will be useful for finding post-transcriptionally controlled genes.

SRS suit allows users to get information by the cross-references between TRSIG entries if some homologous elements are found. The TRSIG site at RIKEN (http:/www.rtc.riken.go.jp/jouhou/trsig/trsig.html) provides detailed tutorial and examples of entries. Pilot version of the TRSIG database compiles the information on various translational signals of higher plants. Currently, it contains data on 16 translational signals, 74 experiments, 119 sequences of regulatory regions, and 114 reconstructed full-sized sequences. We plan to enlarge the TRSIG content and supplement it with post-transcriptional signals of vertebrates.

Acknowledgments

Authors are grateful to Maria Goretti (RIKEN) for the design of TRSIG site at RIKEN. This work was supported by the RIKEN grant for bilateral investigations and STA grant (to A.Kochetov). It was also supported in part by the Russian Foundation for Basic Research (02-04-48508) and Program for Support of Scientific Schools (00-15-97968).

References

1. Kochetov A.V., Grigorovich D.A., Titov I.I., Vorobiev D.G., Syrnik O.A., Vishnevskii O.V., Sarai A., Kolchanov N.A. (2001). mRNA-FAST (mRNA-Function, Activity, STructure) computer system. Mol. Biol. (Mosk.). 35: 1039-1047.
2. Niepel M., Gallie D.R. (1999). Identification and characterization of the functional elements within the tobacco etch virus 5' leader required for cap-independent translation. J. Virol. 73:9080-9088.

COMPARATIVE COMPUTATIONAL ANALYSIS OF 5'-REGION OF CYTOPLASMIC TYROSYL-TRNA SYNTHETASE GENE IN HIGHER EUKARYOTES

Nazarenko M.M., Odynets K.A., * Kornelyuk A.I.

Institute of Molecular Biology and Genetics, National Academy of Sciences of Ukraine, Kyiv, Ukraine,
e-mail: kornelyuk@imbg.org.ua
*Corresponding author

Key words: *orthological genes, housekeeping genes, transcription factor binding site, promoter model and module, CpG islands, Scaffold/Matrix Attachment Regions*

Resume

Motivation: Comparative analysis of 5'-regions of orthological housekeeping genes by computational tools allows estimating quickly their conservative regulatory pattern and revealing the differences that have emerged during evolution. *Results:* The structure of 5'-region of TyrRS gene was examined in four eukaryotic organisms in various aspects. In general it shows the typical features of housekeeping genes. Orthological sequences lack TATA box in appropriate position. In the case of human and mouse they are located within CpG islands and include Sp1 sites, what is expected for vertebrates. The approximate boundaries of the promoter zone have been defined for each 5'-region. Conservative promoter framework has been determined for orthological TyrRS genes by computational generating of four promoter models. One common hexameric palindrome has been predicted. One Scaffold/Matrix Attachment Region has been found in 5'-region of mouse gene. This work is a striking instance of testing the bioinformatics tools on the specific biological objects.

Introduction

The comprehensive study of the 5'-region of eukaryotic genes is of a great importance in functional gene analysis. 5'-region of eukaryotic gene encompasses the first exon beginning and regulatory structures located either in transcribed or untranscribed gene sequence. The promoter zone, represented by a number of various transcription factor binding sites (TFS), must be there necessarily. Scaffold/Matrix Attachment Regions (S/MAR), by which chromatin DNA is anchored on nuclear matrix, is also expected in 5'-region, because such physical contact is required for the initiation of transcription. It was demonstrated that a range of crucial nuclear processes, such as replication and recombination, are initiated at S/MAR as well (Travers, 1994; Bode et al., 1992). Virtually, each S/MAR is a complex of various consensuses, such as Ori pattern, topoisomerase II sites, AT-richness, kinked DNA patterns. Palindromic motifs can be also expected in the 5'-region, as they form hairpins or loops that facilitate protein binding, active in the sites of transcription, replication or recombination.

In our study we have performed thorough online computational analysis of the 5'-region of cytoplasmic tyrosyl-tRNA synthetase (TyrRS) gene in four eukaryotes: human (*Homo sapiens*), mouse (*Mus musculus*), fruit fly (*Drosophila melanogaster*) and nematode (*Caenorhabditis elegans*). The major purpose was to learn, how well its structure corresponds to the expected for eukaryotic housekeeping genes and what regulatory features are the most conservative in orthological sequences.

Methods and Algorithms

Genome sequences used for TyrRS gene analysis. All the sequences were found from the NCBI GenBank database. *H. sapiens*: AL356459, chromosome 1, contig NT_004511.8, HTGS_PHASE1; *M. musculus*: AL607123, Chr. 4, HTGS_PHASE1; *D. melanogaster*: AE003527 genomic scaffold 142000013386050, Chr. 3L; HTG (HTGS_PHASE3). Gene is completely annotated as CG4561; *C. elegans*: AL132880, cosmid Y105E8E, Chr. I; HTG (HTGS_PHASE3). Gene is annotated as Y105E8E.v.

Databases of eukaryotic promoters: *Eukaryotic Promoter Database* v.70 (http://cmgm.stanford.edu/help/manual/databases/epd.html). Programs for promoter prediction: *Genomatix PromoterInspector* (http://genomatix.gsf.de/cgi-bin/promoter inspector/promoterinspector.pl); *BDGP Neural Network Promoter Prediction* (http://www.fruitfly.org/seq_tools/promoter.html); *Promoter recognition program* (http://wwwmgs.bionet.nsc.ru/mgs/programs/recon2/); *Markov Chain Promoter Finder McPromoter* v3.0 (http://promoter.informatik.uni-erlangen.de/); *WWW Promoter Scan* (http://bimas.dcrt.nih.gov/molbio/proscan/); *CGD Nucleotide sequence analysis (TSSG and TSSW)* (http://genomic.sanger.ac.uk/gf/gf.shtml); *Promoter 2.0 Prediction Server* (http://www.cbs.dtu.dk/services/promoter/); *Human Core-Promoter Finder by Michael Zhang* (http://argon.cshl.org/genefinder/CPROMOTER/human.htm). Program for defining promoter models: the task "Extract common framework" in the category "Analyze your sequences" from the manager *GEMS Launcher* (http://www.genomatix. de/cgi-bin/gems/launch.pl). Program for transcription factor binding sites prediction: *Genomatix*

MatInspector Professional. Program for TATA box prediction: *WebGene HCtata Hamming Clustering Method* for TATA Signal Prediction in Eukaryotic Genes (http://l25.itba.mi.cnr.it/~webgene/wwwHC_tata.html). Program for CpG islands prediction: *WebGene* CpG islands prediction (http://l25.itba.mi.cnr.it/genebin/wwwcpg. pl). Program for the prediction of S/MAR and S/MAR consensuses: *MAR-Wiz* (http://www.futuresoft.org/MAR-Wiz/). Program for palindrome search: *TRES* (http://bioportal.bic.nus.edu.sg/tres/).

Implementation and Results

Nucleotide sequences of human and mouse orthological TyrRS genes were obtained by genome BLAST alignment with the respective TyrRS cDNA. As we hadn't found any fly and nematode cDNA, annotated as TyrRS cDNA, we had to submit human cDNA for their TyrRS gene search, expecting their high homology. As a result, we found High Throughput Genomic Sequences (HTGS) of the first stage of progress for human and mouse genes, and of the third stage for fly and nematode.

Then we calculated the most probable transcription start site (TSS) for each orthological TyrRS gene. It was found as the average beginning of all expression sequence tags (EST) well aligned. In this study we didn't trust the longest EST variants, considering them as the products of alternative splicing or background transcription. For the next analysis we tried to cut out the equal genomic regions, so that calculated potential TSS were at the distance from 200 to 250 bp from the end of such fragments and corresponded to the sense strand of orthological TyrRS genes. Consequently, DNA sequences 1351, 1350, 1420 and 1430 bp long were obtained for human, mouse, fly and nematode respectively and then analyzed. Defining the boundaries of the potential promoter, we applied the original approach. Predictions of different search programs were overlapped for the sequence of each organism, and the overlapping of at least three program predictions, adjacent to the potential TSS, was considered as a potential promoter. We always restricted the search by the promoter library of correspondent organism and applied default cutoffs. Calculated promoter zones were positioned relatively to potential TSS. As it is shown in the Table 1, found promoter regions are at the large distances from TSS, which contradicts to the idea of minimal promoter in the close vicinity to TSS. However, we think these regions should significantly overlap with real promoters.

Table 1. Defining the boundaries of the potential promoter zones.

Organism	Number of used programs	Overlapping length, bp	Distance to average TSS, bp	Maximal number of overlappings achieved within the region of three overlappings
H. sapiens	9	328	+97	6
M. musculus	9	334	+168	5
D. melanogaster	7	267	-88	4
C. elegans	6	185	+450	4

To reveal conservative regulatory patterns in orthological promoter sequences we applied the strategy of framework search in a set of sequences (Werner, 2000). As a result, four 3-component promoter models were generated computationally. Promoter models are combinations of TFS with definite order and relative distances. In the case of analyzed sequences, these patterns are common for human, mouse, fly and nematode 5'-regions. Specificity of the promoter models (framework score – FS-Score) ranged from 0.44 to 0.88. The model with the highest FS-Score is represented in Table 2.

Table 2. 3-component model with the highest FW-score generated by *Genomatix ModelGenerator*. Matrix similarity reflects the probability of each element. Distance ranges are caused by different variants-matches of this promoter model.

№	Element	Strand	Matrix similarity	Distance to next element	FW-Score
1.	Human and murine ETS1 Factor	±	Optimized (min. 0.87)	25–86 bp	0.80 / 0.80
2.	E2F-myc activator/cell cycle regulator	±	Optimized (min. 0.78)	11–73 bp	
3.	Signal transducer and activator of transcript. Factor	±	Optimized (min. 0.73)	–	

The set of promoter models was used to scan Eukaryotic Promoter Database with input cutoff 0.95 and resulted by 13 matches. Interestingly, those six matches were annotated as adenovirus promoters. Independently, we compared the analyzed human, mouse, fly and nematode sequences to the database of promoter modules, which are conservative pair combinations of TFS. Although no module, common for all four sequences, was obtained, three modules were common for human and mouse: ETSF_ETSF_01; SP1F_SP1F_01 and SP1F_CEBP_01. Notably, there were overall 3 copies of Sp1 site in human modules and 7 copies in mouse modules. It is well known that Sp1 is a typical TFS for promoters of housekeeping genes in vertebrates. (Mudge et al., 1998).

Analysis of C+G content resulted by characterization of human and mouse sequences as CpG islands with such properties: C+G content was 67 and 68%; CpG content was 7 and 8% in human and mouse respectively. Along with the copies of Sp1 site this is expected for the promoters of housekeeping genes in vertebrates (Pedersen, 2001).

The orthological 5'-regions were then submitted to TATA box search by specialized program. Besides, the sequences were analyzed for the presence of all TFS. None of these programs demonstrated appropriately positioned (about +30 bp) TATA box, which was expected for the promoters of eukaryotic housekeeping genes. However, two other conservative proximal promoter elements – CCAAT box, GC box – were located in one copy within the region [–200- +200] bp in human and mouse sequences, but in (–)-orientation. While searching the most complex common palindromes we found the single hexameric structure GGANNNNTCC with such positions relatively to the calculated TSS: -32 bp in human; -142 bp/mouse; -810 bp/fly; -275 bp/nematode.

At last, we characterized S/MAR consensuses in orthological sequences. Fly and nematode differ by the larger number of potential Ori patterns (15 and 17, respectively) in comparison to human and mouse (3 and 1, respectively), and nematode sequence is AT-enriched (161 matches in contrast to 15 ones in fly and no matches in others). There were 3 matches of topoisomerase II site in human; 1 in mouse; 4 in fly and 6 in nematode. Kinked DNA was found 2 times in human and mouse, 4 in fly and 2 in nematode. Based on S/MAR features the program *MAR-Wiz* predicted potential S/MAR region, but we succeeded only with mouse sequence. Predicted S/MAR was 201 bp long and positioned at -382 bp relatively to the potential TSS.

Discussion

With the wide range of programs, we have characterized such features of 5'-regions of orthological TyrRS genes as promoters, palindromes, CpG islands and S/MAR patterns in comparative manner and concluded about their functional correlation. One S/MAR was predicted in mouse 5'-region, but all analyzed sequences had a number of other S/MAR features: topoisomerase II sites, Ori pattern, AT-richness, kinked DNA. Orthological promoters have demonstrated typical properties of the promoters of housekeeping genes, which lack TATA box and in the case of vertebrates contain CpG islands and Sp1 sites. However, CCAAT and GC boxes were located close to TSS in human and mouse sequences, what is unusual for eukaryotic housekeeping genes. Despite the evolutionary divergence, these 5'-regions have retained the conservative framework of transcriptional regulation, represented in our study by 4 promoter models. The latter have proved to be close to adenovirus promoters, what may elucidate common patterns of their regulation. Since the analysis was based on the computational tools, our results indicate the potential structures of the 5'-regions and need experimental support. The composition of promoter models also needs deeper analysis.

References

1. Bode J., Kohwi Y., Dickinson L., Joh T., Klehr D., Mielke C., Kohwi-Shigematsu T. (1992) Biological significance of unwinding capability of nuclear matrix-associating DNAs. Science. 255, 195-197.
2. Mudge S., Williams J., Eyre H. et al. (1998) Complex organization of the 5`-end of the human glycine tRNA synthetase gene. Gene. 209, 49.
3. Pedersen A., Baldi P., Chauvin Y., Brunak S. (1999) The biology of eukaryotic promoter prediction – a review. Comput. Chem. 23, 191-207.
4. Travers A. (1994) Chromatin and transcription. DNA-Protein interactions. 7, 166-171.
5. Werner T. (2000) Target gene identification from expression array data by promoter analysis. Biomol. Eng. 17, 87-94.

COMPUTER ANALYSIS OF mRNA UNTRANSLATED REGIONS OF HYPOXIA-INDUCED CORN GENES

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

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: titov@bionet.nsc.ru
[2] RIKEN Institute, Tsukuba, Ibaraki 305-0074, Japan
*Corresponding author

Key words: *computer analysis, mRNA, translation, 5'UTRs, 3'UTRs, secondary structure*

Resume

Motivation: Untranslated regions (UTRs) of mRNA frequently house determinants of translational activity. UTR variability and ambiguous localizations of these signals in the sequence in combination with their intricate compositions, including elements of primary and secondary structures, underlie the difficulties arising while detecting translation determinants. Searching for such signals requires a combination of contextual analysis techniques with search for invariant elements of the secondary structure.

Results: Two groups were distinguished among the analyzed mRNA UTRs of hypoxia-induced corn (*Zea mays*) genes. Typical of the first group are longer 5'UTRs with a pronounced secondary structure that brings into spatial proximity the cap site and site of translation initiation. 3'UTRs of this group contain more AUG triplets compared with the second group. 5'UTRs of the second group are shorter and lack a stable secondary structure. Presumably, an intergroup distinction between mechanisms of translation regulation underlies this segregation of mRNAs with respect to the contextual and structural characteristics.

Introduction

Signals affecting mRNA translational activity could play an important role in gene expression control. The majority of cellular 5'UTRs are scanned by ribosomes from the cap site to the site of translation initiation. Length of the leader sequence, false translation initiation sites, secondary structures, or proteins binding to this region may modulate this process. 3'UTRs also contain quite a few determinants of translational efficiency; however, mechanisms of their actions are yet vague. Thus, translation determinants may be present over the entire length of UTRs, whose length amounts from several dozens to several thousands nucleotides, may be degenerate, and depend strongly on one another. These signals have complex structure, are highly diverged, and dispersed within sequences, challenging conventional bioinformatics methods to reveal them. A successful approach to the problem may consist in developing computer methods that combine analysis of sequence homology and alignments with secondary structure calculations and its comparison. We present an example of such an integrated analysis for revealing mRNA-located translational signals.

Methods

UTRs of hypoxia-induced genes were extracted from database on translational enhancers (Kochetov et al., 2001; http://wwwmgs.bionet.nsc.ru/mgs/dbases/trsig/). To localize the candidate mRNA regions capable of forming non-randomly stable structures and to refine the secondary structure translational signals, we used the algorithm GArna (Vorobiev et al., 2002; http://wwwmgs.bionet.nsc.ru/Programs/2dStructRNA/). The final evolutionary invariant model was constructed using the program GenBee (http://www.genebee.msu.su/genebee.html).

Results and Discussion

5'UTRs. The sample of 5'UTRs comprised six sequences. These sequences distinctly fell into two groups with reference to the stability of their secondary structure and their lengths: the first group appeared to form stable structure, while second did not (Table). Moreover, stability of each sequence belonging to the second group was inferior compared to the typical stability for the sequence of the same nucleotide composition (see Z-score value). Interestingly, all the mRNA 5'UTRs studied so far by different authors (see Titov et al., 2002 for review) displayed the secondary structure that was at least equal in its stability to that of the shuffled random sequences. It is likely that the regions from the second group had undergone a selection directed against stability of the secondary structure.

Further search for invariant secondary structure involved only the three sequences of the first group. The sequences in question were aligned with additional weights ascribed to invariant complementary pairs. Consequently, a model of the secondary structure common for the entire group was constructed (Fig.).

Table. Sequence length and secondary structure energy of hypoxia-induced mRNA leaders.

	mzesus1b	mzeadh1cm	zmadh2n	zmaldoar	zmenola	zmsucs2
Length, nt	123	107	126	74	54	71
Energy, kcal	–19.3	–19.5	–16.8	0	0	0
Z-score	–0.98	–0.59	0.57	0.33	0.62	0.63

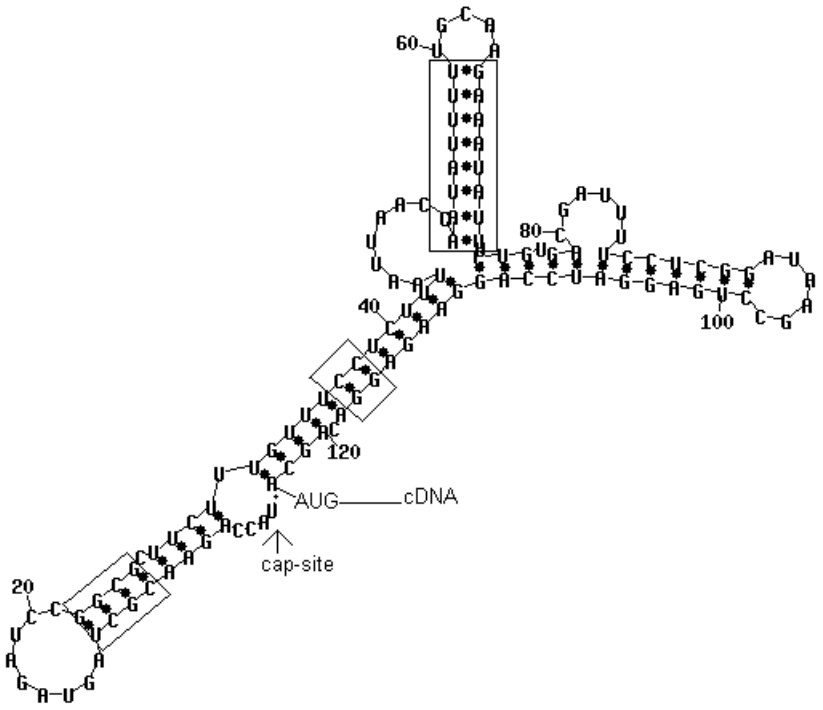


Fig. Secondary structure model of 5`UTR of mzesus1b (Table 1). The helices supported by comparative analysis are shown in rectangles.

An interesting specific feature of this structure is a spatial proximity of the cap site and the site of translation initiation. This proximity may be involved in a interdependent regulation of binding of translation factors with these sites. This regulation may involve both the inhibition (for example, via a simple shielding of one site when binding to the other) and stimulation provided by an additional stabilizing interaction between two factors bound to mRNA. Elucidation of these interactions requires experimental study.

3'UTRs. The sample of 3'UTRs comprised 11 sequences. The total number of AUG triplets in the sample, 30, appeared approximately equal to the expected value amounting to 31.1, calculated on the assumption on independence of the neighboring positions. However, the six genes whose 5'UTRs were considered above fell into the same distinct groups of three sequences in each (Table): the first group contained three AUG triplets (versus 27 expected randomly), whereas the second only 20 (versus expected 25).

While both the excess or shortage of AUG triplets in 5'UTRs are comparatively well studied (Bernardi, 2000; Kochetov et al., 1998), their increased content in 3'UTRs were detected only in RNAs of certain viruses (Hann et al., 1997). Presumably, they are involved in a short translation re-initiation in 3'UTRs, thereby retaining ribosomes on mRNA and enhancing their efficient transfer to the translation start.

Thus, the mRNAs of corn genes studied allowed us to detect a group of mRNAs distinguishable by the following contextual and structural characteristics. Their 5'UTRs exhibit the secondary structure that brings the cap site with the site of translation initiation and their 3'UTRs contain increased number of AUG triplets. Presumably, simultaneous occurrence of these two specific features represents a signal indicating the presence of the translation determinants facilitating the rotation of ribosomes on the mRNA.

Acknowledgements

The work of IIT and AVK was performed while their visit to RIKEN with a support of STA Foundation.

References

1. Bernardi G. (2000). The compositional evolution of vertebrate genomes. Gene. 259:31-43.
2. Kochetov A.V., Grigorovich D.A., Titov I.I., Vorobiev D.G., Syrnik O.A., Vishnevskii O.V., Sarai A., Kolchanov N.A. (2001). mRNA-FAST (mRNA-Function, Activity, STructure) computer system. Mol. Biol. (Mosk.). 35:1039-1047.
3. Vorobiev D.G., Titov I.I., Ivanisenko V.A. (2002). GArna Internet resource for the analysis of the RNA secondary structure: its status in 2002, Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

4. Titov I.I., Vorobiev D.G., Ivanisenko V.A., Kolchanov N.A. (2002). A fast genetic algorithm for RNA secondary structure analysis. Chem. Bull., in press.
5. Hann L.E., Webb A.C., Cal J.-M., Gehrke L. (1997). Mol. Cell. Biol. 17(4):2005-2013.
6. Kochetov A.V., Ischenko I.V., Vorobiev D.G., Kel A.E., Babenko V.N., Kisselev L.L., Kolchanov N.A. (1998). FEBS Lett. 440:351-355.

EFFECTS OF CORRELATIONS DURING RIBOSOME MOVEMENT ALONG mRNA

[1] Titov I.I., [2] Sarai A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: titov@bionet.nsc.ru
[2] RIKEN Institute, Tsukuba, Ibaraki 305-0074, Japan
*Corresponding author

Key words: *correlation, statistical mechanics, mRNA, translation, initiation, termination, codon*

Resume

Motivation: During translation, ribosomes move linearly along mRNA. In this process, a simultaneous presence of two ribosomes at one codon is impossible; however, this fact is disregarded in all the known models of translation. In other words, the concentration of ribosomes on mRNA is conventionally considered small, that is, the translation is assumed inefficient.

Results: The problem of stationary movement of the ribosome pool of the length L along the mRNA consisting of N codons is solved exactly by methods of statistical mechanics for $1<<L<<N$. Mutual correlations of the ribosome movement yield three qualitatively distinct phases. Realization of a phase depends on what process—initiation, termination, or elongation—controls the translation process; hence, it is feasible to detect the limiting phase from the pattern of ribosome distribution along mRNA. The model proposed can be easily restated for any case of a linear transport of ribosomes (for example, their scanning of 5'UTRs). Dispersion of the ribosome codon reading speeds is considered for the case of shock translation.

Introduction

During translation, a ribosome is moving along mRNA reading one codon after another. The rate of ribosome codon reading is controlled by a large number of factors and processes: the content of a necessary tRNA, conformational transitions in the ribosomal complex, mRNA structure etc. However, another factor of not least importance yet avoided the consideration in analytical models of translation: a ribosome covers few tens of codons and cannot overlap or pass one another during a linear scanning. This mutual impenetrability yields nonvanishing correlations in their mutual arrangement. Taking this "trolleybus effect" into account is a compicated multiparticle problem.

Movement of mutually impenetrable particles was considered in a large number of physical phenomena (Richards, 1978; Haus, Kehr, 1987; Bouchaud, Georges, 1990; Titov, Yakobson, 1991)—diffusion and chemical reactions in solids, car parking and traffic, etc. It was found that excluded volume of particles results in considerable correlations between occupations n_i and n_j of the neighboring positions (Richards, 1978). It is essential that calculation of the pair correlator $<n_i n_j>$ requires calculation of the triple correlators, etc., leading to an infinite hierarchy of equations. As a rule, a "naive" simplification of the problem by uncoupling the correlator $\langle n_i n_j \rangle \approx \langle n_i \rangle \langle n_j \rangle$ is justified only in the range of vanishingly small n_i. This means that the conventional equations of physico-chemical kinetics are applicable in these problems only to description of the limit case of small particle concentrations.

Two cases are known—a one-dimensional lattice and a medium strongly nonuniform in the hopping rates—when the effect of correlations is amplified to such a degree that not only the effective characteristics of the problem (diffusion coefficients, reaction rates, etc.) are renormalized, but also the transport laws itself are changed (Richards, 1978; Haus, Kehr, 1987; Bouchaud, Georges, 1990; Titov, Yakobson, 1991). For example, the value of a formally defined diffusion coefficient appears equaling zero. As the ribosomes are incapable of passing one another while moving along mRNA (a "trolleybus effect") and inequivalence of codons is well known, both these situations are typical of translation. Consequently, neglecting of the correlation effects is justified only in the case of inefficient translation (a single ribosome).

In this work, we are constructing a statistical mechanics model of translation. It appears that three qualitatively distinct modes of translation are feasible depending on the rates of the three main processes involved (initiation, elongation, and termination). The least number of ribosomes are loaded onto mRNA in the case when translation is controlled by initiation. In this mode, the kinetics—ribosome movement and relaxation of heterogeneity in their location—is fastest. The most densely populated mRNA is found in the case of limitation of the process by termination, accompanied by slowest kinetics. The control of elongation maximizes the total rate of protein synthesis. Thus, the pattern of ribosome arrangement on mRNA may provide the information on the stage that is limiting the translation. A consequence of this model is the hysteresis in the transition from one mode into another, manifesting itself in expression oscillations. Interestingly, there is a point whereat all the three modes coexist; a system being in the neighborhood of this point is most

flexibly regulable. A transparent restatement of the problem demonstrates that the effect observed is typical of any linear ribosome movement, the scanning of 5'UTRs by pre-translational complex included.

Theory

Let us consider a set of identical mRNAs limiting the consideration to their coding regions comprising N codons. Let ribosome cover an mRNA region with a length of L codons. Let us first determine the rates of elementary processes disregarding a screening effect of one ribosome on another and characterize these rates with a number of ribosomes involved in the process per unit time summed over all the mRNAs. Then, the translation initiation rate K_i is the number of ribosomes that enter the first codon. Similarly, the termination rate K_t the number of ribosomes leaving the last codon. The rate of peptide chain elongation K_0 is characterized by a probability of a ribosome to shift by one codon (let this rate be equal for all the codons). Let also the mRNA ensemble to be in a steady state, that is, to be characterized by a ribosome flux along mRNA that is constant in the time domain. Although the ensemble of molecules is non-equilibrium, its time invariance allows the methods of statistical mechanics to be applied.

In the situation of one ribosome per one mRNA, all the characteristics of translation are trivially calculable using the above parameters. However, if the ribosome density is finite, they may represent mutual hindrances of their movement. For example, the probability n_i of the front of a ribosome to be located at the ith codon depends on the occupation numbers n_{i+L} and n_{i+L+1} of another ribosome (taking into account that the length of ribosome is L):

$$\frac{d<n_i>}{dt} = <n_{i-1}>(1-<n_{i+L}>) - <n_i>(1-<n_{i+L+1}>)$$

where square brackets indicate the averaging over time history.

Thus, the evolution of single occupations $<n_i>$ depends on pairwise correlations $<n_i n_j>$. The time dependence of the pairwise correlations is expressed via triple etc. correlators, leading to an infinite hierarchy of linked equations. Hence, the problem can be solved only for time invariant $<n_i>$. Let us brief the sequence of operations omitting cumbersome algebraic calculations.

Let the statistical weight of a certain ribosome configuration on mRNA be described as

$$W(n_1, n_2, ..., n_N) = \frac{1}{Z}\left\langle L \left| \prod_{i\{L\}=1}^{N} (n_i P + Q(1-n_i)) \right| R \right\rangle \quad (1)$$

where Z is the partition function; the product takes into account the size of ribosome L; P and Q, are matrices of transfer from one codon to another; and L and R, left and right state vectors. It is essential that equation (1) can be simplified from the limitations on the size L in the product by using the L-time bigger matrices. This is a trick which let us to generalize Derrida's solution of a driven lattice gas problem which corresponds to our particular case L=1 (Derrida, Evans, 1997). Using the boundary conditions and the relations between occupations (descending from the fact that a ribosome covers L codons simultaneously), let us calculate the matrix elements.

At the last step using the weights (1), the average occupation $<n_i>$ and correlator $<n_i n_j>$ are calculated (the latter is indispensable for calculating the rate of relaxation of ribosome density fluctuations).

Results and Discussion

For the sake of simplicity, let us right away restrict our attention to the asymptotic of biological $N>>L>>1$ interest at a fixed density of ribosomes loaded onto mRNA, $n=<n_i>=\text{const}$. Table 1 summarizes the following averaged characteristics of the translation model:

$n,=<n_i>/L$ density; measured as a specific (per one mRNA) number of ribosomes per the number of codons N;

$V=K_0(1-nL)$, the speed of movement; measured as a specific number of codons traveled by the ribosome per unit time;

D, Brownian diffusion coefficient; corresponds to the following experiment: we "marked" all the ribosomes located at a specified codon at a time moment $t = 0$ (for example radiolabeling the corresponding amino acid); then, we are monitoring as the package of labeled ribosomes is diffusing with time t. The package dispersion is proportional to t with a proportionality coefficient D.

$J=nV$, the total rate of protein synthesis; determined as a number of codons traveled by all the ribosomes together per unit time. This value is not independent and is determined by the first two characteristics.

The calculations yield that one of three solution types is possible for each set of rates $\{K_i, K_t, K_0\}$ (Table).

Table. Three states of translation.

Control type	$min(K_i, K_t) > K_0/2$; elongation control	$min(K_0/2, K_t) > K_i$; initiation control	$min(K_0/2, K_i) > K_t$; termination control
Density, n	Mediocre, $1/2L$	Low, K_i/LK_0	High, $(1-K_t/K_0)/L$
Speed, V	Mediocre	Fast	Slow

Relaxation rate, D	Mediocre	Fast	Slow
Total synthesis rate, J	Maximal	$K_i(1-K_i/K_0)$	$K_t(1-K_t/K_0)$

Thus, depending on the ratio of speeds, the ribosomes form one of the three possible phases. The phase exhibiting the maximal translation efficiency is realized when elongation plays the role of translation limiting factor. In the case of translation controlled by initiation, the phase with low ribosome density is formed. Finally, the ribosome density is maximal when the limiting factor of the process is termination. The number of ribosomes loaded onto mRNA is easily countable experimentally, allowing thereby the stage of the process to be detected.

From the biological standpoint, the situation when all the three rates are coordinated, $K_i=K_t=K_0/2$, is most interesting. In the neighborhood of this point, minor changes in one of the rates cause a passage to another phase. The transition between the low and high ribosome density phases is accompanied by a jump in the equilibrium number of ribosomes on mRNA. Consequently, interesting modes of translation efficiency (oscillatory and other) may emerge, additional feedback between ribosome concentration and one of the rates provided. Existence of such a feedback has been experimentally proved for several microorganisms.

The described model of translation is a strong simplification, as it disregards an essential specific feature of actual translation—nonequivalence of the codons, which is a factor involved in regulation of translation efficiency. Account of this factor may renormalize the translation characteristics considered; however, the general pattern will be qualitatively retained. New qualitative features may appear when the first moment of distribution with respect to protein chain elongation times diverges. Models of anomalous diffusion (Bouchaud, Georges, 1990; Titov, Yakobson, 1991) suggest that in this situation, ribosomes will be joined in the packages moving rapidly from one rare codon to another. In this process, the macroscopic speed of their movement will be power-law dependent on time, meaning, in essence, its nontrivial dependence on the length of mRNA.

Apparently, the model proposed may be used for describing scanning of 5'UTRs by pre-translational complexes. Here, melting the secondary structure of the leader would represent a new and additional factor underlying the spatial correlations of ribosomes (more precisely, their clustering).

Acknowledgements

The work of IIT was performed while his visit to RIKEN with a support of STA Foundation.

References

1. Bouchaud J.-P., Georges A. (1990) Anomalous diffusion in disordered media: statistical mechanics, models and physical applications. Phys. Rep. 195(4/5), 127-293.
2. Derrida B., Evans M.R. (1997) V.Privman (Ed.), Non-equilibrium statistical mechanics in one dimension, Cambridge University Press, Cambridge, 277.
3. Haus J.W., Kehr K.W. (1987) Diffusion in regular and disordered lattices. Phys. Rep. 150(5-6), 263-406.
4. Titov I.I., Yakobson B.I. (1991) Concentration-dependent diffusion coefficient within the lattice gas model of disordered crystal. Rev. Solid State Sci. 5(1), 53-59.
5. Richards P.M. (1978) Correlated hopping conductivity in a general two sublattice structure. J. Chem. Phys. 68(5), 2125-2128.

UNOPTIMAL TRANSLATION START SITE CORRELATES WITH INCREASED CONTENT OF IN-FRAME DOWNSTREAM AUG CODONS AT THE BEGINNING OF CDS OF EUKARYOTIC mRNAs

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

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Tsukuba Institute, Institute of Physical Chemical Biology (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074 Japan
*Corresponding author: e-mail: ak@bionet.nsc.ru

***Key words**: translation start site, eukaryotic mRNA, downstream AUG*

Resume

Motivation: It is known that the context of start AUG codon influences the translation initiation efficiency of eukaryotic mRNAs. It is expected that mRNAs would contain the translation start site (TSS) in optimal context to support a high translation rate. However, the fraction of eukaryotic mRNAs with unoptimal TSS was found to be relatively large (Suzuki et al., 2000; Peri, Pandey, 2001). It is likely that some unknown mRNA features compensate for the TSS "weakness" and enhance its recognition.

Results: We performed a comparative analysis of yeast mRNAs with "weak" and "strong" translation initiation sites. We have found that these samples differ in the content of downstream in-frame AUG codons ($dAUG^{in}$): occurrence of "weak" TSS correlates with significantly higher $dAUG^{in}$ content at the beginning of CDS. Presumably, some mRNAs with unoptimal TSS can be translated efficiently due to the presence of nearby downstream AUGs, and this mRNA feature should be taken into account while predicting mRNA translation efficiency.

Introduction

It is known that the recognition of translational start codon by eukaryotic ribosomes depends on its nucleotide context (Kozak, 1999). For mammalian mRNAs, the most crucial elements of AUG codon context are the adenine at position -3 and (less significant) guanine at position +4. It was found that mutations at these positions could decrease considerably the rate of translation from AUG codon. In yeast, mutations in the start codon context lead to a moderate decrease in mRNA translation rate, and its efficiency depends on the nucleotide occupying position –3 upstream of AUG (A>G>C>U; Yun et al., 1996). As was found recently, a large portion of eukaryotic mRNAs annotated in nucleotide sequence databanks contained TSS in an unoptimal context (Suzuki et al., 2000; Peri, Pandey, 2001). We assumed that some unknown mRNA features could in part compensate for the negative effect of unoptimal TSS context. To test this hypothesis, we performed the comparative analysis of yeast genes with "weak" and "strong" TSS. We have found that (1) the frequency of proximal dAUG in-frame with CDS is significantly higher if TSS has a "weak" context and (2) the distributions of in-frame and out-of-frame dAUGs along CDS differ, namely, the occurrence of in-frame dAUGs is considerably higher at the beginning of CDS. Potential heterogeneity of the TSS position and the problems of mapping the genuine TSS(s) are discussed.

Methods

A reliable set of 4128 non-redundant yeast nuclear-encoded ORFs was taken from http://biochimica.unipr.it/wobble. The 500-nucleotide long sequences located upstream of the start codon of 6329 yeast genes were taken from ftp://genome-ftp.stanford.edu/pub/yeast. By combining sequences with the same identifier, we prepared for contextual analysis the sample of fragments spanning the regions between –90 nucleotide upstream of the start AUG and nucleotide 300 of CDS.

TSS was classified into "strong" and "weak" taking into account the nucleotide occupying position –3 upstream of AUG (purine and pyrimidine, respectively; according to Yun et al., 1996).

Results

We analyzed the sample of yeast genes for the occurrence and characteristics of the first downstream AUG triplet within a proximal 300-nt long fragment of the coding sequence (4010 of 4113 mRNA contain AUG triplet(s) in this region; see Table). We found that approximately 20% of the yeast genes contained "weak" TSS (i.e., pyrimidine in position –3 upstream of AUG). We compared two mRNA subsamples with the translational start codons in either "weak" or "strong" context and took into account the position of dAUG with respect to TSS and its reading frame (in-frame or out-of-frame with TSS). It was found that some statistical features of the downstream AUG triplet correlated with the TSS "strength". First, the mRNAs with a "weak" TSS contained the first proximal dAUG in-frame with CDS significantly more frequently than the mRNAs with "strong" TSS: 31% versus 22%; this difference was statistically significant according to both T-test

for independent samples (t_{value} = –5.66, $P<10^{-7}$) and the Mann–Whitney U-test (Z = 4.11, $P<10^{-4}$). It may mean that 5'-dAUG^{in} is positively selected and may be of functional importance to compensate for the TSS "weakness".

Table. Characteristics of 5'-proximal AUG triplets* within 300-nucleotide long 5'-CDS fragments.

Genes	With "strong" TSS (3203)	With "weak" TSS (807)
5'-dAUG^{in}	701	251
5'-dAUG^{out}	2502	556

*5'-dAUG^{in}, 5'-proximal AUG triplet in-frame with CDS; 5'-dAUG^{out}, 5'-proximal AUG triplet out-of-frame with CDS.

The distributions of in-frame and out-of-frame AUGs along 90-nucleotide long fragments downstream the start codon of yeast genes with a "weak" TSS is shown in Fig. It is evident that the distributions of dAUG^{in} and dAUG^{out} are different (Fig.).

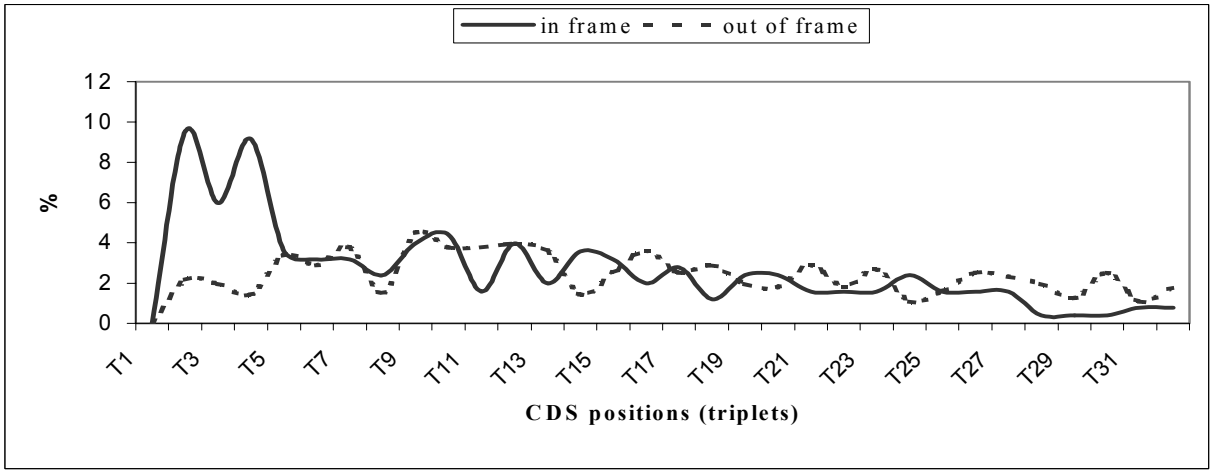


Fig. Distribution (%) of the proximal downstream AUG triplets either in-frame or out-of-frame with TSS within 90 nucleotide long CDS 5'-end fragment (in-frame T1 is the start AUG codon). Only mRNAs with unoptimal translation start codon were selected (807 sequences, see Table 1). Frequencies of the out-of-frame AUGs (+1 and +2) are summarized.

Discussion

It is well known that recognition of translational start codon by eukaryotic ribosomes in the most cases depends on its nucleotide context (Kozak, 1999). mRNAs with a "weak" TSS produce polypeptides less efficiently: 40S ribosomal subunits can miss TSS in a "weak" context and initiate translation at the downstream AUG codon by a leaky scanning mechanism (Yun et al., 1996). If 5'-dAUG lies out-of-frame with CDS and starts translation of a short internal ORF, the premature termination could result in rapid mRNA degradation by nonsense-mediated decay (NMD) and the average level of cytoplasmic stability of such mRNA will be low (Welch, Jacobson, 1999). Thus, a "weak" TSS is a strong negative feature, and eukaryotic genes could evolve some characteristics to compensate for the TSS "weakness".

In this work, we analyzed sequence features of 4113 non-redundant genes of *S. cerevisiae* (putative proteins were excluded). It appeared that 20% of the yeast genes in this sample contained pyrimidine at –3 position upstream of the start AUG codon and, thereby, can either produce additional protein(s) or be subjected to nonsense-mediated decay if some other mechanisms would not prevent it. It can be expected that closely located downstream AUG codons in-frame with TSS may compensate for the TSS "weakness". If such dAUG^{in} occurs in an optimal context, it can eliminate the negative influence of unoptimal translation initiation signal. Even if dAUG occurs in suboptimal or unoptimal context, it could enlarge the portion of efficiently translated mRNAs by avoiding scanthrough and RNA degradation by NMD mechanism.

We analyzed the AUGs that could be met by 40S ribosomal subunits scanning beyond a "weak" TSS. We found that the distributions of dAUG^{in} and dAUG^{out} along CDS differed considerably (Fig.): e.g., 62 of 251 mRNAs contained 5'-dAUG^{in} at CDS positions 2 to 4, whereas only 31 of 556 mRNAs contained 5'-dAUG^{out} at these positions. It may mean that a more frequent occurrence of 5'-dAUG^{in} in an optimal context can be of functional importance. These data argue in favor of the hypothesis that evaluation of the real TSS "strength" should take into account not only AUG codon context but some additional factors. Statistical analysis of TSS (e.g., Peri, Pandey, 2001) should take into consideration at least the nearby downstream in-frame AUG codons, since it is likely that N-truncated polypeptides starting from close downstream in-frame AUGs have the same functions as their full versions and, thereby, these dAUG^{in} can be considered as either additional TSS increasing polypeptide synthesis rate or genuine TSS where most translation events occur.

Acknowledgments

This work was supported by the RIKEN grant for bilateral investigations, STA grant (to A. Kochetov), Russian Foundation for Basic Research (grant № 02-04-48508), and the Program for Support of Scientific Schools (grant № 00-15-97968).

References

1. Kozak M. (1999). Initiation of translation in procaryotes and eukaryotes. Gene. 234:187-208.
2. Peri S., Pandey A. (2001). A reassessment of the translation initiation codon in vertebrates. Trends Genet. 17:685-687.
3. Suzuki Y., Ishihara D., Sasaki M., Makagawa H., Hata H., Tsunoda T., Watanabe M., Komatsu T., Ota T., Isogai T., Suyama A., Sugano S. (2000). Statistical analysis of the 5'untranslated region of human mRNA using "oligo-capped" cDNA libraries. Genomics. 64:286-297.
4. Welch E.M., Jacobson A. (1999). An internal open reading frame triggers nonsense-mediated decay of the yeast SPT10 mRNA. EMBO J. 18:6134-6145.
5. Yun D-F, Laz T.M., Clements J.M., Sherman F. (1996). mRNA sequences influencing translation and the selection of AUG initiator codons in the yeast *Saccharomyces cerevisiae.* Mol. Microbiol. 19:1225-1239.

STRUCTURAL FEATURES OF MRNA REGION AT THE TRANSLATION START SITE

**** Likhoshvai V.A., Kochetov A.V., Matushkin Yu.G., Kolchanov N.A.***

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

Key words*: translational signals, translation start site (TSS), secondary structure, expression efficiency*

Resume

Motivation: It is known that recognition of the translation start site (TSS) depends on the contextual features of neighboring RNA regions. However, many eukaryotic mRNAs stored in databanks contain the start AUG codon in unoptimal context. We assumed that structural mRNA features could also be involved in TSS recognition to increase mRNA translation initiation efficiency.

Results: Stability of the predicted secondary structure was found to be lower within 5'UTR compared with coding sequence (CDS). Notably, we found a significant negative correlation between the predicted secondary structure stability (evaluated with LCI index) and predicted expression level (evaluated by synonymous codon bias, EEI index) in the TSS region. We assume that stability of the secondary structure in the region of translation start site can be an important factor in TSS recognition.

Introduction

It is well known that the secondary structure of eukaryotic mRNAs is an essential factor influencing the translation rate (Kozak, 1999). The mRNA functional domains differ in the sensitivity to secondary structure (SS): it has been shown that the negative effect of moderately stable hairpins in 5'-untranslated region (5'UTR) is very strong, whereas only very stable SS in the mRNA coding part can influence translation rate. It has been found that different 5'UTR subdomains differ in the sensitivity to secondary structure: it has been shown that 5'-terminal part of the leader is more sensitive to SS (i.e., hairpins of lower stability can decrease translation) than the other subdomains in vertebrate cells (Kozak, 1999; Niepel et al., 1999). Another important feature influencing the mRNA translation is the context of translational start codon (Yun et al., 1996). It was found that unoptimal context (pyrimidine in –3 position upstream of AUG) decreased translation initiation and mRNA translation efficiencies. However, analysis of nucleotide sequence databanks has shown that a large part of eukaryotic mRNAs is characterized by poor context of the start codon (Suzuki et al., 2000; Peri, Pandey, 2001). We assumed that mRNA structural features could influence the TSS recognition. To test this hypothesis, we analyzed structural characteristics of eukaryotic mRNAs in the region of translation start site.

Materials and Methods

Sequence data. (1) *S. cerevisiae* coding DNA sequences (CDS) with 600-nucleotide-long 5'- and 3'-end extensions were extracted from the GenBank using the Feature Table information. (2) The sample of yeast full-sized mRNAs was extracted from the EMBL databank (http://www.embl-heidelberg.de/). Full-sized 5'UTRs were selected from the entries containing description of mapped transcription start sites and complete coding regions. This resulted in a set of mRNA 5'UTRs of 171 non-redundant (<60% identity with CDS) yeast genes. Finally, we used in analysis 98 of 171 genes characterized by a single transcription start site.

Methods

The procedure used for calculating the elongation efficiency index (EEI) of the protein coding regions (CDS) and the procedure for ordering the genes are detailed in (Likhoshvai, Matushkin, 2002). LCI for the *j*th nucleotide of an *i*th fixed sequence was calculated using the following equation:

$$LCI(i,j)=\sum_{\substack{m:\ m\le j\le m+s-1\ \ or\\ m+s+l-1\le j\le 2m+2s+l-2}} 10\left\{\sum_{s=s_{\min}}^{s_{\max}}\left[\sum_{l=l_{\min}}^{l_{\max}}-\psi\left(con(m,\ m+s-1),\overline{con(m+s+l-1,2m+2s+l-2)}\right)\right]\right\}, \qquad (1)$$

where $c1 = con(m,m + s - 1)$, the gene context from mth to $(m + s - 1)$th nucleotide and $c2=\overline{con(m+s+l-1,\ 2m+2s+l-2)}$, the complementary inverted gene context from $(m + s + l - 1)$th to $(2m + 2s + l - 2)$th nucleotides. If the words $c1$ and $c2$ are identical, then $\psi(c1, c2)$ is the energy of the local secondary structure where $con(m, m + s - 1)$ and $con(m, s + l - 1, 2m + 2s - l - 2)$ form the stem, while the sequence fragment between them is the loop. If $c1 \neq c2$ *or* $\psi(c1,c2)\big|_{c1=c2} \ge o$, we assume that $\psi(c1,c2) = 0$. The length of accountable

inverted repeat is not less than s_{min} and not more than s_{max}; the distance between accountable inverted repeats, not less than l_{min} and not more than l_{max}. The energy of secondary structures was calculated conventionally (Turner, Sugimoto, 1988).

The codon adaptation index (Sharp, Li, 1987) was calculated using the program CodonW 1.3 (written by J.Peden and available at http://www.molbiol.ox.ac.uk/cu). The mRNA base-pairing probabilities (McCaskill, 1990) were calculated using Vienna RNA secondary structure package (http://www.tbi.univie.ac.at/~ivo/RNA/). Pearson's product–moment correlation coefficient (r_p) was used for the measurement of correlation between two variables.

Results

I. *Analysis of 5' gene regions.* The LCI profiles along 5'UTR (300 nucleotides upstream and downstream of AUG) are demonstrated in the Fig. 1a, c. It is evident that LCI declines from about –200 nt to the start AUG codon and increases from +1 to +50 nt. To verify the functional significance of the LCI distribution, we estimated the correlations between LCI in this region and the expression level evaluated by EEI index (taking into account the codon bias, Likhoshvai and Matushkin, 2002). We divided all the genes into 12 groups according to increase in this index. The first group comprised 500 genes with the lowest values of EEI index. The 500 genes displaying next levels of the EEI values composed the second group, and so on. The twelfth group contained 304 genes exhibiting the maximal indices (the result was insensitive to the partition method used). Then, we discarded from these groups all the genes containing less than 400 nt in their open reading frames and calculated the mean EEI values and averaged LCI profiles for the remaining genes of all the groups.

At the next stage, we calculated the Pearson's correlation coefficients between the vector of EEI values and the vector of LCI values for each nucleotide and calculated the significance of the correlations found using Student's test. The longest continuous region displaying a significant negative correlation ($p < 0.01$) insensitive to the parameters used for calculating LCI was the [–19; +13] fragment of 5'UTR (Fig. a, c). In the case the significance level was decreased to 98%, this fragment expanded to [–20; +25]. In Fig., vertical lines indicate the borders of this fragment.

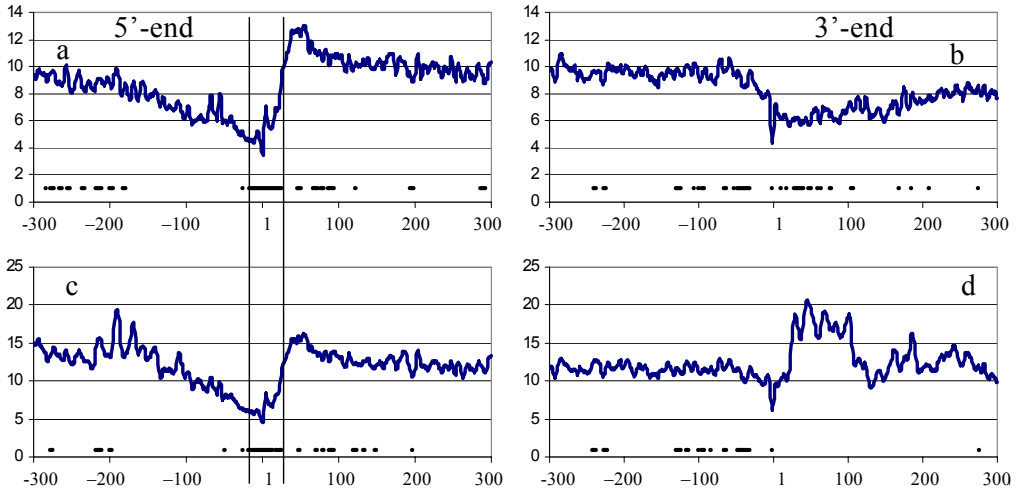


Fig. LCI-profiles of 5'- and 3'-regions of CDS: (a, c) 5'-regions; (b, d) 3'-regions; the abscissa, distances in nt relative to the initiation and stop codons [+1, +3]; the ordinate, values of LCI; bold intermittent line, positions with significant correlations ($p<0.01$) between LCI and EEI values; (a, b) calculations taking into account perfect hairpins with stems of 3–6 nt and loops of 3–50 nt; and (c, d) stems of 3–12 and loops of 3–38 nt.

II. *Analysis of 3'-region.* It may be expected that secondary structure at the 3' CDS border (i.e., near to stop codon) can also influence the mRNA translation. To check this assumption, we calculated LCI profiles along the 3'-mRNA ends. We found that the region [–55; +1] also displayed a decrease in LCI, although less pronounced (Fig. b, d). The LCI profile of the region to the right of the stop codon depends on the parameters used in calculation. When long hairpins (with a stem of up to 12 nt) were taken into account, a region of increased averaged LCI values was detected (Fig. 1d). When the hairpins with a stem of up to 6 nt were taken into account, the region with a high averaged LCI values was absent (Fig. b). Analysis of correlations between LCI profile and EEI index showed no continuous significant correlations. Moreover, within the 3'UTR or CDS fragments (other than 5'-end neighbors), the significant correlation stretches were relatively short and depended on the parameters used for calculating LCI (compare Fig. a with 1c and 1b with 1d, respectively).

The LCI value in the protein coding regions displays a trend of positive correlation with the expression level. However, it is possible that this trend to a considerable degree is determined by codon selection. Numerical experiments on generation of random sequences with a certain codon composition confirm this hypothesis (data not shown).

Discussion

It has been found experimentally that the secondary structure decreases the translation activity of 5'UTR by slowing the ribosome movement, whereas no strict evidence on the negative influence of CDS- or 3'UTR-located hairpins on mRNA expression have been found (Kozak, 1999; Niepel et al., 1999). We analyzed the distribution of LCI index, reflecting the potential base pairing of nucleotides along the borders of CDS with 5'UTRs and 3'UTRs. We assumed that the secondary structure could represent an additional signal marking the transition between the coding and 5'-untranslated parts of yeast mRNA. We found that there was a marked shift in LCI index at the border of 5'UTR and CDS. This difference in potential secondary structure can represent an additional signal increasing the efficiency of AUG recognition. Moreover, there is a significant negative correlation between LCI and EEI values, possibly reflecting selection for a more optimal TSS in highly expressed mRNAs.

Table. Correlation between base-pairing probabilities at positions around the start AUG codon and the codon adaptation index.

To verify the correlations described above, we examined an additional sample compiled from the reliable yeast mRNAs with the mapped transcription start site. To evaluate the secondary structure and expression level, we also used the independent criteria: codon adaptation index (CAI; Sharp, Lee, 1987) instead of EEI and base-pairing probabilities (BPP; McCaskill, 1990) instead of LCI. Correlation coefficients between the BPP for the fragment from –10 nt upstream of AUG codon to 10 nt downstream and CAI are listed in the Table. It is evident from Table 1 that there are significant negative correlations between BPP and CAI at the majority of

Positions	r_p	Positions	r_p
–10	–0.14	A	–0.06
–9	**–0.23**	T	0.04
–8	**–0.30**	G	–0.11
–7	**–0.23**	+4	0.00
–6	**–0.24**	+5	–0.01
–5	**–0.23**	+6	0.11
–4	**–0.20**	+7	–0.01
–3	**–0.22**	+8	–0.05
–2	**–0.26**	+9	–0.14
–1	–0.13	+10	–0.05

STUDY OF THE RELATIONS BETWEEN EXPRESSION LEVEL AND CONTEXTUAL CHARACTERISTICS OF YEAST GENE FUNCTIONAL REGIONS BY THE ZET METHOD

[1] Pichueva A.G., [2] Kochetov A.V., [1] Zagoruiko N.G.,*

[1] Institute of Mathematics, SB RAS, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: anna_@math.nsc.ru
[*]Corresponding author

Key words: *contextual characteristics, gene regions, expression level, prediction strategy, algorithm for filling empiric tables (ZET)*

Abstract

Motivation: It is known that certain contextual characteristics of extended functional regions of genes influence their expression (Kochetov et al., 1998); however, the mechanisms underlying this phenomenon yet require studying. Reported here are the results obtained by computer analysis of dependences between the levels of expression and contextual characteristics of various yeast gene functional regions (5'UTR, promoter, CDS, and 3'UTR).

Results: An algorithm allowing the expression level of a gene to be predicted from characteristics of its functional regions, the predicted activities of different functional regions to be compared, and an integrated evaluation of the expression level to be obtained is proposed in this work. It has been demonstrated that characteristics of noncoding regions and expression parameters are interrelated, suggesting that these characteristics are significant for maintaining a high level of expression.

Introduction

The gene expression pattern is determined by a variety of signals localized to functional regions (promoters, introns, exons, and mRNA) and controlling mRNA transcription, splicing, translation, and stability in the cytoplasm. These signals (such as TATA box, transcription factor binding sites, splicing sites, translation enhancers, AUUUA elements, etc.), represented by local regions within genes with a mean length of 5 to 15 bp, are now actively studied. However, numerous experimental data demonstrate that the entire extended functional region structure is also an essential factor affecting the level of gene activity while its expression (Kochetov et al., 1998). These characteristics yet require further studies; however, their significance for predicting the level of expression is undoubtedly high. The main goal of this work was to detect the patterns relating the contextual characteristics of gene regions to the level of expression.

Yeast genes with the known transcription start (Kochetov et al., 2000) were used as the sample of nucleotide sequences. This allowed us to form subsamples of four gene functional regions, namely, mRNA 5'-untranslated region (5'UTR), mRNA 3'-untranslated region (3'UTR), basal promoter (PROM), and coding sequence (CDS). Codon adaptation index (CAI) was used as a criterion reflecting the level of gene expression; frequencies of mono- and dinucleotides (the ratio of observed to expected frequencies), as contextual characteristics of functional regions.

Methods: ZET Algorithm

The ZET algorithm (Zagoruiko et al., 1986; Zagoruiko, 1999) is designed to predict the values of missing elements in a table (the mode of filling gaps) and to edit (verify) the overall table or its parts (the mode of editing). Typical of the actual data tables is their excessiveness, as many properties (columns) are interrelated with one another by certain dependences. The ZET algorithm allows such relations and similarities to be discovered and used to predict the values sought for.

Operation of the ZET algorithm comprises three stages. At the first stage, a "competent" submatrix is selected for a given empty cell *b(ij)* from the initial "object–property" matrix, where the columns are normalized according to dispersion. For this purpose, *t* strings (objects) that are most similar to the *i*th string are initially selected. Then, *t* columns (properties) most pronouncedly correlated with the *j*th column are selected for these strings. At the second stage, the parameters of an equation used for predicting the missing element that would minimize the prediction error are automatically defined. Finally, the element in question is predicted using this equation at the third stage.

The algorithm exploits a linear dependence both between the strings and between the columns of the table. The linear regression equation for the element *b(ij)* is used to calculate the "prompts" basing on the columns *b(k)* and strings *b(l)* of the competent submatrix (Fig. 1) to average these prompts with the weights proportional to the degree of "competence" *L* of the columns and strings in question. The competence is calculated as a function of similarity *r* and mutual completeness of the columns (strings) *p*: $L = r \times p$.

	x(j)		x(...)		x(k)		x(n)
	CAI	///	A_LDR	///	E_SCORE	///	TT5_OE
MISCCO1A	0.16	...	30.00	...	0.76	...	1.13
S61567	0.16	...	36.62	...	0.70	...	1.00
///	...	...	...	...	...	...	...
SCACT	**b(ij)**	...	33.33	...	**b(ik)**	...	1.48
///	...	...	...	...	...	...	...
SCHAP	0.09	...	31.82	...	0.81	...	0.00
///	...	...	...	...	...	...	...
SSCARG56	**b(lj)**	...	46.94	...	**b(lk)**	...	0.62

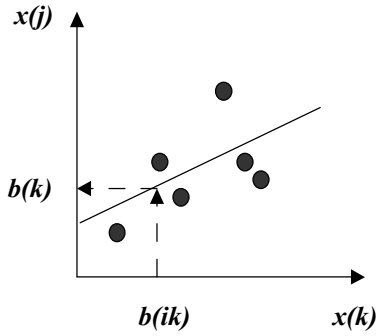


Fig. 1. Filling in the missing element b(ij).

For example, using the "prompts" basing on columns, we obtain the predicted value *b(j),* resulting from the excessiveness of data contained in *t* columns:

$$b(j) = \sum b(k) L^{\alpha}(jk) \Big/ \sum L^{\alpha}(jk), \quad k = \overline{1,t},$$

where α is the coefficient controlling the effect of competence on the result of prediction. The difference in competences is weakly pronounced for small α values, whereas more competent columns exert higher effects for large α values. To choose the parameter α, all the known elements of *j*th column are predicted at different values of this parameter with subsequent choosing of the α value providing the minimal prediction error. The value $\delta(j)$ is further considered as an estimate of the expected error of column-based gap filling. The actual error is considered equal to

$$d = [b(ij) - b^{*}(ij)] / \bar{b}(j),$$

where *b(ij)* is the actual value; $b^{*}(ij)$, predicted value; and $\bar{b}(j)$, mean of the values from *j*th column (hence, the actual error may exceed 100%).

The procedure of filling a gap using the dependence between the *i*th string and all the rest strings selected is similar to the procedure described for columns. Here, all the known elements of the *i*th string are used to make the selection at the minimal value of the error of their prediction $\delta(i)$. The final prediction *b(ij)* of the missing element value is produced by selecting either the prediction based on columns *b(j),* if $\delta(j) < \delta(i)$, or the prediction based on strings, if $\delta(i) < \delta(j)$. The predictions may be also averaged using their weights that are inversely proportional to the value of expected error. The expected error may also be obtained from dispersion of "prompts".

The program ZET outputs the information on what contextual characteristics were included into the competent submatrix. This allows the particular characteristics that are most informative for each individual gene to be revealed. In addition, the frequency of the presence of a characteristic in competent submatrices reflects a mean informativeness of this characteristic.

Description of Experiments: Results and Discussion

The program ZET was run in "editing" mode. The size of competent submatrix was specified as 3*3. The expected errors of predicting CAI values and the CAI values predicted by the program were obtained for 171 yeast genes. Predictions were performed for each functional region separately (Table 1).

How may the final prediction value be obtained from the four prediction variants? To answer this question, different strategies were studied and compared with respect to the least mean value of actual prediction error (39.69) and the highest correlation (0.69) between the predicted and true values of the parameter CAI. According to these criteria, the strategy implying averaging of the three best predictions with reference to the expected error (Figs. 2, 3) was selected. The averaging here is made using the corresponding values of expected errors as weight coefficients:

$$\sum_{i=1}^{3} (b^{i}_{\min} / \delta^{i}_{\min}) \Big/ \sum_{i=1}^{3} (1 / \delta^{i}_{\min}),$$

where δ^{i}_{min} is the first, second, and third minimal values of the expected error for the gene in question (obtained form its different functional parts); b^{i}_{min}, the corresponding predicted values (Table 1).

Table 1. The procedure for obtaining the final prediction value.

		5'UTR (5)		CDS (C)		PROM (P)		3'UTR (3)		Best prediction strategy		
Gene number	Value of CAI marker	Expected error, %	Predicted value	Expected error, %	Predicted value	Expected error, %	Predicted value	Expected error, %	Predicted value	Gene fragments selected as "expert" regions with minimal expected error, % (in ascending order)	Final normalized predicted value	Actual deviation, %
	Mean	*13.17*		*7.20*		*9.73*		*12.03*				*39.69*
1	0.16	1.90	0.17	7.41	0.18	3.58	0.15	12.01	0.62	5 (1.90), P (3.58), C (7.41)	0.17	4.06
2	0.16	9.70	0.45	0.81	0.15	8.99	0.26			C (0.81), P (8.99), 5 (9.70)	0.18	5.91
3	0.12	9.81	0.17	4.32	0.27	16.10	0.27			C (4.32), 5 (9.81), P (16.10)	0.24	44.78
///	///	///	///	///	///	///	///	///	///	///	///	///

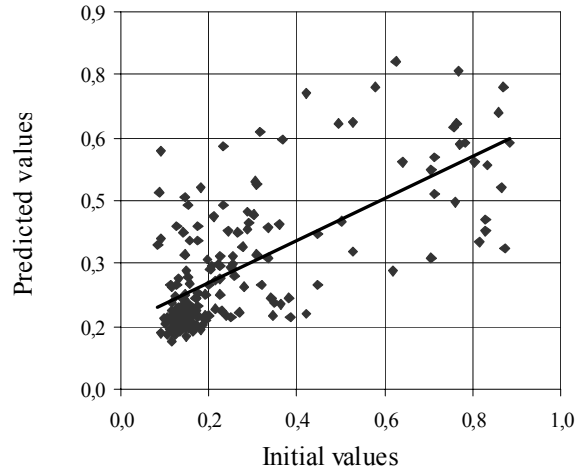


Fig. 2. Dependence of predicted versus true values of the parameter CAI.

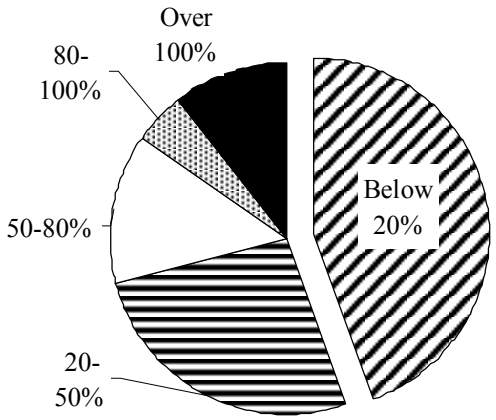


Fig. 3. The number of elements whose prediction errors fall in the corresponding ranges.

Evaluation of Informativeness

The number of competent submatrices generated by ZET algorithm wherein an individual characteristic from each group is present reflects the relative informativeness of the characteristic. For example, G content (5.65%); length (5.46%); contents of TG (5.07%), GA (4.87%), GC (4.87%), and GT (4.87%) dinucleotides, etc., are informative characteristics of the 5'UTR functional regions (Fig. 4). However, additional experiments based on other approaches are necessary for clarifying the relations between the significant characteristics. The informativeness of each of the four gene regions may be also evaluated according to the number of genes for which a region in question gave prediction with a minimal expected error. For example, of 171 genes studied, the minimal expected error resulted from 5'UTR characteristics of 33 genes (0.19); CDS, 87 genes (0.51); PROM, 43 (0.25); and 3'UTR, 8 (0.05). However, note that the 3'UTR subsample was of the smallest size. Taking into account the number of instances with the minimal actual error for each region, we obtain the following rates: 5'UTR, 38 (0.22); CDS, 93 (0.54); PROM, 33 (0.19), and 3'UTR, 7 (0.04). Thus, characteristics of CDS regions appeared the best in both cases. However, the rest gene functional regions may also be used for predicting the level of gene expression.

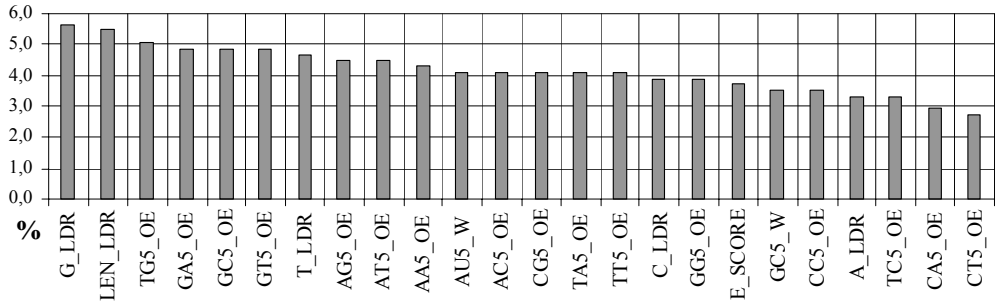


Fig. 4. Informativeness of 5'UTR characteristics.

Conclusion

The experiments performed demonstrate that many genes display significant dependences between contextual characteristics of the four functional regions in question and the level of gene expression. These dependences allow the level of expression to be predicted with an accuracy differing for individual groups of genes. It is possible to separate groups of genes with a high or a medium degree of such dependences as well as indicate the genes whose expression is very weakly influenced by their contextual characteristics. These results suggest that the dependence of expression level on the context is not uniform and allow the relative significance of each functional region to be evaluated.

Acknowledgements

The work was partially funded by the Russian Foundation for Basic Research (grants № 01-07-90376, 02-04-48508, 02-07-90355, 00-04-49229, and 02-01-00082); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Kochetov A.V., Ischenko I.V., Vorobiev D.G., Kel A.E., Babenko V.N., Kisselev L.L., Kolchanov N.A. (1998). Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. FEBS Lett. 440:351-355.
2. Kochetov A.V., Vorobiev D.G., Sirnik O.A., Kisselev L.L., Kolchanov N.A. (2000). Contextual features of yeast mRNA 5'UTRs potentially important for their translational activity. Proc. 2nd International Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk. 1:67-70.
3. Zagoruiko N.G., Elkina V.N., Emel'yanov S.V., Lbov G.S. (1986). OTEKS Applied Software (for Data Analysis). M.: Finansy i Statistika.
4. Zagoruiko N.G. (1999). Applied Methods for Data and Knowledge Analyses. Novosibirsk: Izdatel'stvo IM.

SHORT-RANGE CORRELATIONS IN GENE EXPRESSION PROFILES

[1] Titov I.I., [2] Pal'yanov A.Yu.*

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: titov@bionet.nsc.ru
[2] Novosibirsk State University, Novosibirsk, Russia
*Corresponding author

***Key words**: highly expressed genes, statistical analysis, mRNA, autocorrelation function*

Resume

Motivation: It was demonstrated that the expression level of genes of a number of organisms is related to their affiliation with GC-rich isochors forming a kind of "genome core" (Bernardi, 2000). It was first hypothesized for mammalian genes (Titov et al., 2000) and then discovered while performing computer analysis of the human transcription map that highly expressed genes are clustered within chromosomes (Caron et al., 2001), although the cluster sizes were not evaluated.

Results: We analyzed statistically the expression profiles of human chromosomes 6 and 21 (Caron et al., 2001). In both cases, the autocorrelation function is exponential, which is normal for short-range correlations. This range determines a typical size of the cluster of genes with a similar expression level. The size evaluated appeared amounting to approximately half chromosome length. Roughly speaking, human chromosomes 6 and 21 consist of two similar gene blocks, each displaying its own expression level. The mechanism underlying emergence of these correlations is discussed.

Introduction

Many eukaryotic genomes contain the so-called isochors, that is, extended DNA regions with a relatively constant GC content (Bernardi, 2000). It is assumed that isochores form a "genome core" in the genomes of certain organisms, this core consisting of genes with high expression level (Bernardi, 2000). Our analysis of contextual and structural characteristics of mRNAs of high- and low-expressed mammalian and dicot plant genes has demonstrated a qualitatively distinct behavior of these two taxa (Titov et al., 2000; 2002). We interpreted the results obtained by hypothesizing that the genes displaying similar expression levels were clustered in mammalian chromosomes whereas dispersed in dicot plants (Titov et al., 2000; 2002). (To avoid an apparent contradiction to the results obtained by Bernardi et al., note that they ascribed genomic regions to isochore basing on GC content of the third codon position.)

A present progress in EST analysis allows location of genes on the chromosome to be identified and their expression level to be assessed. Human chromosomes were recently analyzed by this method (Caron et al., 2001), and their results confirmed our hypothesis. This analysis discovered a trend of genes with high expression level to cluster and demonstrated nonrandomness of their localization, although the size of clusters was not calculated. In this study, a more solid statistical technique allowed this size to be evaluated for two human chromosomes.

Data Analyzed

Expression profiles of human chromosomes 6 and 21 averaged over 39 genes were taken from (Caron et al., 2001). These profiles were constructed by computer analysis of EST data through localizing the genes and assessing their expression levels. Expression levels were estimated in number of tags per gene. For the quantitative analysis, we redigitized these profiles.

Method of correlation functions

This method is frequently used for detecting organization of complex objects. Let us consider a series of random variables (l_k). For the sake of simplicity, let us assume that $\langle l_k \rangle = 0$, while the series itself represents a stationary process, so that the correlation function $C(n) = \langle l_k l_{k+n} \rangle$ depends only on n (we also assume that $C(n)$ displays no asymptotic oscillations). Two distinct situations are possible here.

(a) Short-range correlations

In this case, the sum $\sum_{n=1}^{N} C(n)$ converges at $N \to \infty$, that is $C(n)$ decreases with a higher rate than n^{-1} in a limit of large n.

In the case of short-range correlations, the series (l_k) can be partitioned into successive blocks of N_{ident} variables taking similar values. In this process, the values l_k typical of each block will be statistically independent of one another. The block size is estimated by the integral of the following correlation function (Bouchaud, Georges, 1990):

$$N_{ident} = \frac{\sum_{n=1}^{N} C(n)}{\langle l^2 \rangle - \langle l \rangle^2}.$$

Note the most important particular cases of short-range correlations: $C(n) = 0$ everywhere except for $n = 0$ (for instance, an ideal polymer chain) and $C(n) = exp(-n/R)$. In a qualitative manner, the exponential correlation function signals that the variable l_k "remembers" its value at a distance R. For example, this dependence is observed when analyzing a polymer chain with a persistent length R or while simulating a symbol-based sequence with a finite order Markov chain.

(b) Long-range correlations

In this situation, the correlation function decreases as n^{-1} or at a slower rate. This corresponds to a fast increase of the (l_k) fluctuations with N. Such correlations suggest that the sequence (l_k) descends from a dynamic process. They arise when considering a ballistic diffusion (Bouchaud, Georges, 1990) or analyzing musical, linguistic, or genetic texts; heartbeat intervals; or a number of economic indicators (for bibliography, see http://linkage.rockefeller.edu/wli/dna_corr/).

Analysis of expression profiles of human chromosomes

Using the digitized expression profiles (see above), we constructed the correlation functions of expression levels for human chromosomes 6 and 21 (Fig.). Due to finite size of chromosomes, it is reasonable to analyze the behavior of correlation function at distances smaller than 0.1 chromosome length. Within this range, the correlation functions $C(n)$ of expression profiles are well described by an exponential decay ($R^2 = 0.995$). The law of decay allows the typical cluster size with a similar expression level to be calculated, followed by estimating the number of such clusters in a chromosome. Interestingly, despite a considerable difference between the expression profiles of the chromosomes in question, the effective numbers of blocks with similar expression levels are very close, namely, 1.99 and 2.33 for the chromosomes 6 and 21, respectively. Thus, both chromosomes are divided approximately into halves with reference to the expression level.

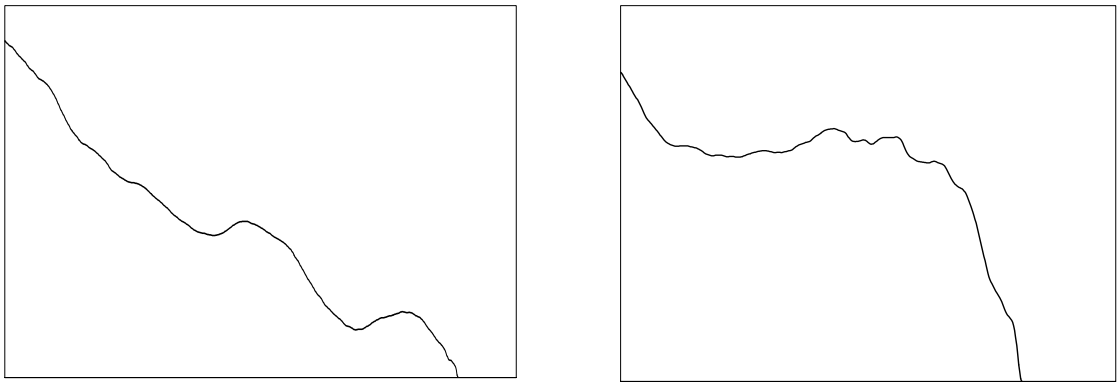


Fig. Autocorrelation functions of the expression profiles of the genes housed in human chromosomes 6 (left) and 21 (right).

Discussion

Exponentially decaying correlations are ascribed to the short-range correlations due to their rapid disappearance. In the case of expression profiles of the chromosomes studied, this term, although terminologically correct, yet fails to reflect the fact that the correlations observed cover a half of the chromosome, or better to say, the entire range of expression level is described by two typical values, each characterizing its half of the chromosome. Analysis of other human chromosomes and comparison of the correlation ranges with chromosome lengths will clarify to what degree this situation of a "chromosome trigger" is typical. A preliminary analysis of human chromosome 11 (data not shown) confirms exponential decay of correlation function, but the correlation radius is much shorter and spans about 0.1 of the chromosome.

Another natural question arises in this connection: what is primary—position of a gene or its expression level? The data available so far do not enable us to select either alternative. The first mechanism of clustering may base on an accelerated recombination between regions related in their nucleotide compositions (and, consequently, in their expression levels—see above on the connection of gene expression to its GC composition). (If genes are uniformly duplicated over the genome, the power-law correlations are to be observed.) Maintenance of a certain specified polymorphism by selection may

represent here an additional factor. On the other hand, an opposite scenario is possible, when expression of a set of clustered genes is controlled via chromatin packaging.

Acknowledgements

The authors are grateful to N.A.Kolchanov for helpful discussions.

References

1. Bernardi G. (2000). The compositional evolution of vertebrate genomes. Gene. 259:31-43.
2. Bouchaud J.-P., Georges A. (1990) Anomalous diffusion in disordered media: statistical mechanics, models and physical applications. Phys. Rep. 195(4/5), 127-293.
3. Caron H. et al. (2001) The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains. Science. 291:1289-1292.
4. Titov I.I., Vorobiev D.G., Kolchanov N.A. (2000) Mass analysis of RNA secondary structures using a genetic algorithm. Proc. BGRS-2000 2:138.
5. Titov I.I., Vorobiev D.G., Ivanisenko V.A., Kolchanov N.A. (2002) A fast genetic algorithm for RNA secondary structure analysis. Chem. Bull., in press.

SEARCHING FOR THE ANTISENSE INTERACTIONS BETWEEN 5'UTR OF EUKARYOTIC GENES

**** Vorobiev D.G., Titov I.I., Omelyanchuk N.A.***

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

Key words: *mRNA, secondary structure, antisense interaction, ribonomics*

Resume

Motivation: At present a fast accumulation of facts concerning participation of antisense conponents in the gene expression regulation takes place. In this connection, it seemed to be interesting to examine a hypothesis about the existance of biologically significant antisense interactions between differenet functional regions of eukaryotic mRNAs.

Results: In this work we present the preliminary analysis of the possible antisense interactions between 5'UTRs of three organisms: *Drosophila melanogaster, Caenorhabditis elegans* and *Arabidopsis thaliana*.

Availability: the program is available from the authors.

Introduction

In the last years knowledge about the translational regulation of the gene expression extends very fast. A class of small RNAs taking part in gene expression regulation has been described for both eukaryots (Lee et al., 2001) and prokaryots (Masse E., et.al., 2002). Such a regulation is carried out by the translation block due to the formation of stable complexes between small RNA and target mRNA molecules. Because of the small size and low homology between each other, it is unclear how many regulatory RNAs are still not identified in sequenced genomes. Therefore, the discovery of several more classes of structural non-coding RNAs in the future is proposed (see for review Doudna, 1999). The existence of antisesense interactions between non-coding regions of mRNAs also can not be excluded, which is confirmed by several recent works (Lehner et al., 2002). In present study we examined this possibility.

We developed an algorithm for quick search in a sample of RNA sequences for the cross-complementarities capable of forming high stable secondary structure complexes. The complexity of the algorithm block that calculates energy of possible complex for given pair of seuences is quadratic to the product of sequences lenghes. We applied the algorithm to search for potential interactions between 5'UTRs of genes in three organisms: *D. melanogaster, C. elegans* and *A. thaliana.*

Methods and Algorithms

Assume that we have a sample of N RNA sequences. To test all possible interactions it is necessary to run over N(N-1)/2 pairs. The calculation of energy for a given pair is itself quite slow. Furthermore, in our task there is a need to determine the relative stability of the potential complex, which requires massive computation with random sequences. For the fastest dynamic algorithms, the time complexity is proportional to $(L_1+L_2)^{2.5}$, where L_1 and L_2 - lengths of sequences. We decided to use for the searching of potential complexes the rough, but fast, algorithm based on the idea of the steepest descent. Let us describe it shortly.

1) Calculate all possible stems for a given pair of sequences S_1 и S_2. (This stage makes algorithm complexity quadratic to the product of sequence lengths L_1L_2). Incomplete helices, whose combinations give rise to “running loops”, are allowed. Energy is calculated for each stem (Jaeger et al., 1989).

2) Add stem giving the best energy gain. Calculate free energy of the formed complex.

3) Check if the stability of the complex improved. If yes, return to step 2. If no, finish energy calculation.

4) Perform the calculation according steps 1-3 for the random sequences produced from S_1 and S_2 by a nucleotide shuffling.

5) Calculate value

$$Z-score(E)=\frac{E_{natural}-<E_{random}>}{\sqrt{disp(E_{random})}},$$

where $E_{natural}$ – complex energy for the pair of sequences S_1 and S_2, E_{random} – same value for the pair of randomized sequences, DE_{random} – it's variance. Negative value of Z-score indicates increased complex stability compared to random sequences.

6) For the cases when Z-score value is lower than prescribed threshold t (<0), calculate complex energy and its Z-score using more precise method implementing secondary structure prediction by genetic algorithm (Vorobiev et al., this volume).

In spite of relative roughness of the described approach, it allows to select for the detailed analysis those pairs of sequences which are capable of forming high stable complexes.

Implementation and Results

According to described method, we performed a search of potential interactions between 5'UTRs (with the lengths varying from 30 to 100 nt) in *C. elegans, D. melanogaster,* and *A. thaliana.* The sequences were extracted from the EMBL data base. On the stage 6 of the protocol pairs with a Z-score<-2 were selected (Table). The distribution of a Z-score statistics turned out to follow the dependence:

$$p(Z) = a\exp(-bZ),$$

which is typical to for limit distributions, which Z-score statistics distribution must belong to by its nature. Deviations from a trend line in the area of high absolute values are accounted for by fluctuation due to a low number of representatives. Differing line slope (and hence differing portion of pairs with a Z-score<-2) between organisms may rise from different general organization of its 5'UTRs, for instance, from the diverse dinucleotide composition.

Table 1. Statistics of the 5'UTR sequences investigated.

	C. elegans	*D.melanogaster*	*A. thaliana*
Number of sequences	237	615	1078
Number of complexes with Z-score<-2	2116	12186	51501
Portion of complexes with Z-score<-2, %	7.57	6.45	8.87
Regression line slope for Z-score distribution (parameter b in the regression formula)	-1.38	-1.53	-1.04

Discussion

The existence of only several evolutionary selected antisense interactions between 5'UTR of different mRNAs must result in appearance of specific modes in the area of high absolute values of Z-score(E) statistics distribution. However, as it is evidently from the Fig. 1, these distributions do not have any marked deviation from the exponential distribution, which is characteristic of random sequences. At the same time, stability of some complexes exceeds 40 kkal/mol by the absolute free energy value, at relatively short loops. This is a very stable complex which may make up an irresistible obstacle for the ribosome (Sagliocco et al., 1993). On the other hand, such an RNA-RNA complex could serve as a target of the gene silencing events or RNA interference (see for review ???), which nature is still unclear.

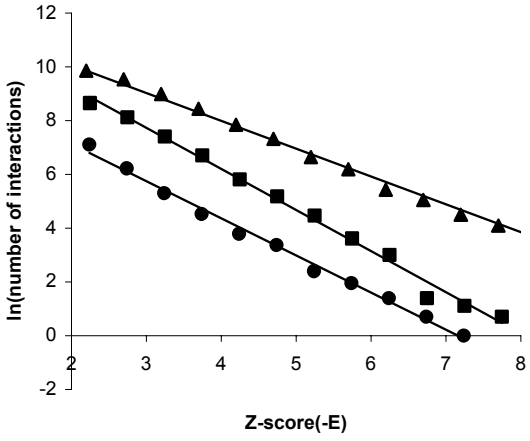


Fig 1. Z-score distribution in semilogarithmic coordinates in the range -∞<Z<-2 for the energy E of complexes formed by all possible pairs of sequences from samples of 5'UTRs in *C. elegans* (O), *D. melanogaster* ( ) и *A. thaliana* (Δ). For convenience absolute values of Z-score are presented.

When searching in a sample of 10^2-10^3 representatives, one may find several dozens of such stable potential complexes. It is clear, that in a sample of 10^3-10^5 RNAs (typical genome size for eukaryotic organism) much more stable complexes could be found. But owing to the existance of mechanisms such as folding RNA into ribonucleoprotein particles, stable complexes does not appear most likely in the cell in general.

On the whole, the question about the biological significance of potential antisense complexes between cellular mRNAs remains undecided and requires more elaborate analysis. Correlation of discovered potential mRNA-mRNA complexes with already known interactions between elements of gene networks seems to be one of the perspective directions for study.

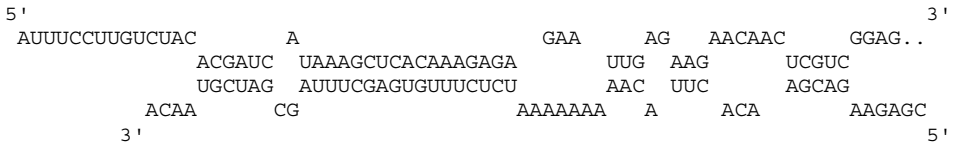


Fig. 2. Secondary structure of the potential complex between 5'UTRs of A. thaliana profilin 1 mRNA (length 100 nt, upper molecule) and glutathione transferase mRNA (length 58 nt). Complex energy is equal to -40.4 kkal/mol, Z-score = -11.6.

If significat mRNA-mRNA interactions exist, they can not be distinguished from the total noise using only complex energy or Z-score. In this case, the specificity of their biological function might be mediated by the context of the concrete gene network.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (grant № 01-07-90376), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), National Institutes of Health USA (№ 2 R01-HG-01539-04A2), The Department of Energy USA (№ 535228 CFDA 81.049). The authors are grateful to N.A.Kolchanov for helpful discussions.

References

1. Doudna J.A. (2000) Structural genomics of RNA. Nat Struct Biol. 7 Suppl., 954-956.
2. Jaeger J.A., Turner D.H., Zuker M. (1989) Improved predictions of secondary structures for RNA. Proc. Natl Acad. Sci. USA. 86, 7706-7710.
3. Lee R.C., Ambros V. (2001) An extensive class of small RNAs in Caenorhabditis elegans. Science. 294, 862-864.
4. Lehner B., Williams G., Campbell R.D., Sanderson Ch.M. (2002) Antisense transcripts in the human genome. Trends Genet. 18, 63-65.
5. Masse E., Gottesman S. (2002) A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. Proc. Natl Acad. Sci. USA. 99(7), 4620-4625.
6. Sagliocco F.A., Vega Laso M.R., Zhu D., Tuite M.F., McCarthy J.E., Brown A.J. (1993) J. Biol. Chem. 268, 26522-26530.
7. Turner D.H., Sugimoto N., Freier S.M. (1988) RNA structure prediction. Ann. Rev. Biophys. Biophys. Chem. 17, 167-192
8. Vorobiev D.G., Titov I.I., Ivanisenko V.A. (2002) GArna Internet resourse for the analysis of the RNA secondary structure: its status in 2002. This volume.

TRANSLATION ELONGATION STAGES CRITICAL FOR THE EFFICIENCY OF GENE EXPRESSION IN UNICELLULAR ORGANISMS

**** Likhoshvai V.A., Matushkin Yu.G.***

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

Key words: *efficiency of gene expression, translation elongation, translation elongation stages, codon usage frequencies, perfect inverted repeats*

Resume

Motivation: Efficiency of gene expression is a basic integral characteristic of gene function. Detection and study of the factors changing the efficiency of gene expression is essential for both a better understanding of the key stages of translation and design of artificial systems with a prespecified expression level.

Results: Patterns of codon compositions and local secondary structures of the protein coding regions in genes of 74 unicellular organisms, whose complete genomic sequences had been determined, were studied. We divided all the microorganisms into the five following groups: (1) the group comprising 32 organisms with the elongation efficiency determined only by the codon composition; (2) the group of 8 organisms with the codon composition inessential for the elongation efficiency; on the contrary, the elongation was limited by the degree of local secondary structures; however, the energies of these local secondary structures played no role; (3) the group consisting of only one organism—*P. aeruginosa* PA01—with no effect of codon composition on elongation efficiency, controlled by both the local secondary structures and their energies; (4) the group of 30 organisms where both the codon composition and local secondary structures were essential for determining the elongation efficiency; however, the energies of the local structures were insignificant; and (5) the group of 3 organisms where all the three factors—codon composition, local secondary structure, and their energies—were essential.

Introduction

Translation of mRNA in unicellular organisms is one of the most energy-consuming stages of the gene expression process. For example, up to 50% material and energy resources may be spent for translation in an *E. coli* cell. Consequently, increase in efficiency of translation machinery operation might play the role of a long-term factor of evolutionary selection. The stage of elongation may be a target of such optimization. The elementary act of elongation—attachment of an amino acid residue to a growing polypeptide chain—comprises three successive stages, namely, placement of the charged isoacceptor tRNA in the ribosome A site, transpeptidation, and translocation. In general case, each of these three elongation stages may be limiting for the polypeptide growth.

In this work, we are relating the efficiency of isoacceptor aminoacyl-tRNA placement with the codon usage frequencies, while the efficiency of translocation, with the number of local secondary structures. Since the modern concepts prevent from relating the transpeptidation stage with the mRNA context directly, we omitted it from consideration in this work. We are determining the relative contributions of codons and local secondary structures to the efficiency of translation elongation for 74 organisms whose complete genomes have been sequenced. This allows us to select most informative characteristics of nucleotide compositions of their protein-coding regions to recognize adequately the efficiency of gene expression.

Materials and Methods

Coding DNA sequences (CDS) were extracted from GenBank using the information contained in the Feature Table.

The average time spent by a ribosome for one elongation act was calculated using the equation $EEI = u_1 T_a + u_2 T_e$. Here, T_a, accurate to the proportionality coefficient, has a meaning of the average time required for isoacceptor aminoacyl-tRNA to be placed in the ribosome A site and is calculated as $T_a(i) = \sum_{j=1}^{n_i} \beta_{\delta(i,j)} \Big/ n_i$, $\beta_\delta = \frac{\sum_{m=1}^{C} \sqrt{\alpha_m}}{\sqrt{\alpha_\delta}}$,

where α_δ has a meaning of the usage frequency of codon δ of the genetic code C within a certain gene subset. The second component $T_e(i)$, accurate to the proportionality coefficient, has a meaning of the average time spent by a ribosome for translocation and is calculated as $T_e(i) = t_{min} \cdot (1 - p(i)) + t_{max} p(i)$, where t_{min} (t_{max}) is the minimal (maximal) conditional time of translocation, respectively, and $p(i)$, the probability of t_{max} realization, calculated according to the equation

$$p(i) = \int_0^{LCI(i)} \frac{k^{n+1}x^n}{G(n+1)} e^{-kx} dx \ , \ k = m/\sigma^2, \ n = (m/\sigma)^2,$$

where m and σ^2, are the expectation and variance, respectively, of a random positive value with a distribution density of $\frac{k^{n+1}x^n}{G(n+1)}e^{-kx}$, where $G(n+1)$ is a gamma function and $LCI(i)$, complementarity index. We used two methods for calculating the complementarity index.

<u>LCI ζ *form.*</u> The averaged number of complementary regions disregarding the energy of secondary structure formation is calculated using equation (2) over a region with a length m_i for a fixed translation frame (here, m_i equals the triple number of codons contained in the *i*th gene plus 55 nucleotides from its 3'-end):

$$LCI\zeta(i) = \frac{\sum_{m=1}^{m_i-2s_{\max}-l_{\max}+1} \left\{ \sum_{s=s_{\min}}^{s_{\max}} \left[\sum_{l=l_{\min}}^{l_{\max}} \zeta\left(con(m,m+s-1), \overline{con(m+s+l-1,2m+2s+l-2)}\right) \right] \right\}}{m_i - 2s_{\max} - l_{\max} + 1}, \quad (2)$$

where $con(i,j)$ is the gene context between *i*th and *j*th nucleotides; $\overline{con(i,j)}$, the corresponding complementary context between *i*th and *j*th nucleotides ($i \le j$); and ζ(conext1,conext2) = 1, if the words conext1 and conext2 are identical, otherwise ζ(conext1,conext2) = 0. The length of accountable inverted repeat falls between s_{min} and s_{max}; the distance between accountable inverted repeats falls between l_{min} and l_{max} (here, $s_{min} = s_{max} = 3$, $l_{min} = 3$, and $l_{max} = 50$).

<u>LCI ψ *form.*</u> In equation (2), ζ is substituted with ψ(conext1,conext2)—the energy of secondary structure potentially formed by a perfect repeat found (here, $s_{min} = 3$, $s_{max} = 6$, $l_{min} = 3$, and $l_{max} = 50$ are used for LCI ψ form). The energies of secondary structures were calculated conventionally (Turner and Sugimoto, 1988).

The following indices were used for calculations: (1) $u_1 > 0$, $u_2 = 0$, when only the codon composition was taken into account; (2) $u_1 = 0$, $u_2 > 0$, regarding only the secondary structure energies; and (3) $u_1 > 0$, $u_2 > 0$ considering both characteristics. For variants (2) and (3), the local complementarities were taken into accounts regarding (ψ from) and disregarding (ζ form) the secondary structure energies. Thus, five variants of EEI (elongation efficiency index) were used in the calculations.

To discover the factors critical for elongation in each particular unicellular organism, we ordered its genes according to the decrease in values of each of the five indices described. Then, we analyzed the order of genes encoding ribosomal proteins in the five produced ordered lists of all the genes of each organism, assuming that these genes belonged to the group with the highest expression. Consequently, the larger is the shift of ribosomal genes from the center of the ordered list of genes, the more reliable is the conclusion on the effect of the characteristic observed (which is taken into account in the corresponding index) on the efficiency of expression. The shifts and their statistical significances were calculated as described in (Likhoshvai, Matushkin, 2002). A separate group of organisms displaying the maximal shift with respect to each particular index variant was formed (Table). The significance of the shift for each organism amounted to >0.999.

<table>
<tr><th colspan="3">Organisms</th><th>Group</th></tr>
<tr><td colspan="3">A. tumefaciens C58, B. halodurans C-125, B. subtilis, B. melitensis, C. muridarum, C. trachomatis, C. pneumoniae, C. pneumoniae AR39, C. pneumoniae J138, E. coli K12, E. coli O157 H7, E. coli O157 H7 EDL933, H. influenzae, L. innocua, L. monocytogenes EGD-e, M. loti, M. leprae, P. multocida, S. typhi, S. typhimurium LT2, S. meliloti, S. aureus Mu50, S. aureus N315, S. pneumoniae R6, S. pneumoniae TIGR4, S. pyogenes, S. pyogenes MGAS8232, Synechocystis sp. PCC6803, V. cholerae, Y. pestis, S. cerevisiae, and S. pombe</td><td>Group A (codon compositions of genes are critical for elongation)</td></tr>
<tr><td colspan="2">B. burgdorferi, Buchnera sp. APS, C. jejuni, H. pylori 26695, H. pylori J99, M. genitalium, M. pulmonis, and U. urealyticum</td><td colspan="2">Group ζ (the amount of local complementarities disregarding the energy are critical for elongation)</td></tr>
<tr><td>P. aeruginosa PA01</td><td colspan="3">Group ψ (the amount of local complementarities and the energy of secondary structures are critical for elongation)</td></tr>
<tr><td colspan="3">A. aeolicus, C. rescentus, C. perfringens, C. acetobutylicum, D. radiodurans R1, F. nucleatum ATCC 25586, L. lactis, M. tuberculosis H37Rv, M. tuberculosis CDC1551, M. tuberculosis H37Rv, M. pneumoniae, Nostoc sp. PCC 7120, R. conorii Malish 7, R. prowazekii, T. maritime, T. pallidum, A. pernix K1, A. fulgidus, Halobacterium sp. NRC-1, M. thermoautotrophicum, M. jannaschii, M. kandleri strain AV19, M. acetivorans strain C2A, P. aerophilum, P. abyssi, P. horikoshii, S. solfataricus, S. tokodaii, T. acidophilum, and T. volcanium</td><td>Group Aζ (the codon composition and the amount of local complementarities disregarding the energy are critical for elongation)</td></tr>
<tr><td colspan="2">N. meningitidis MC58, N. Menin- gitidis Z2491, and X. fastidiosa</td><td colspan="2">Group Aψ (the codon composition and the amount of local complementarities, and the energy are critical for elongation)</td></tr>
</table>

Results and Discussion

The results obtained are listed in table below. Note that none of the principally possible groups is empty. The first group contains 32 organisms; the critical factor for their elongation efficiency is only the codon composition, while secondary structures have no effect on the process. The second group comprises eight organisms, where the translocation step is critical. The translocation rate here decreases with the increase in the degree of local complementarities; however, the secondary structure energies are insignificant. Only one organism—P. *aeruginosa* PA01—forms the third group. Its elongation efficiency is also determined at the translocation stage; however, the energy of local secondary structures is an important factor influencing the rate of ribosome movement. The stage of tRNA attachment is not critical for the organisms belonging to the second and third groups; in this respect, they are opposite to the organisms of the first group. The fourth group contains 30 organisms, the fifth, only three. In these organisms, both elongation stages considered influence the elongation efficiency. However, the elongation efficiency of the organisms from the fourth group, similarly to the second, depends only on the presence of complementary regions, whereas the energy of secondary structures remains insignificant. In the fifth group, as in the third, the elongation efficiency also depends on the energies of secondary structures.

There are certain common patterns evident from the results listed in the below Table.

First, note that the groups A and Aζ are most numerous, comprising together 62 organisms of the 74 analyzed. Uniting them with the organisms of group ζ, we get 70 organisms where the energy of secondary structures fails to limit the elongation process versus only 4 organisms where this energy plays an essential role in determining the elongation efficiency.

Second, all the studied archaebacteria (underlined in the Table) fall into group Aζ. The majority of related eubacterial species also fall into the same groups. The eukaryotic organisms (S. *cerevisiae*, and S. *pombe*) analyzed also belong to one group A. Thus, evolutionary related species display a trend of grouping together.

The results obtained suggest several biological interpretations.

Resulting groups of unicellular organisms

This produces the impression that organisms had used different strategies while optimizing the elongation stage during their evolution. The organisms belonging to the most numerous group A have evolved such efficient mechanisms of mRNA physical movement through the ribosome (or ribosome movement along mRNA) that the translocation stage ceased to be limiting (if it ever was). The selection of these organisms towards increasing the elongation efficiency was achieved through optimizing codon compositions of their genes.

On the contrary, the data obtained suggest that the translocation stage in the organisms belonging to groups ζ and ψ is sensitive to the local hindrances formed, in particular, due to local complementarities. However, the placement of tRNA in the ribosome A site is either equally efficient for all the codons or proceeds in parallel with the process removing the hindrances ahead of the moving ribosome (preparatory stage of translocation) and this process is slower. Thus, the latter process shields in a sense the former process, thereby providing the evolutionary neutrality of the codon mutations along with the evolutionary sensitivity of the mutations decreasing the amount of local complementarities. In the rest groups, both stages influence essentially the elongation efficiency, suggesting consequently that both characteristics—the codon composition and local complementarities—have been optimized.

Groups ζ and Aζ, on the one hand, and groups ψ and Aψ, on the other, suggest that the corresponding organisms might utilize different mechanisms for their ribosomes to overcome the hindrances represented by local secondary structures at the stage of translocation. We may hypothesize that in the case of group ζ and Aζ organisms, encounter of a ribosome with a hindrance triggers a mechanism that spends a predetermined batch of resources (time or energy) to remove all the hindrances within mRNA region of a certain length. On the contrary, the corresponding mechanism of group ψ and Aψ organisms is somehow capable of estimating the hindrance "capacity" and spends the proportional time (or, possibly, energy) for its removal. Note that the total number of organisms displaying the sensitivity of elongation to the secondary structure energies appeared insignificant, amounting to only four.

Thus, we have demonstrated that traces of evolutionary optimization of translation elongation are detected is all the organisms analyzed. These traces are fount at the levels of gene codon compositions, local secondary mRNA structures, or both at once. All the microorganisms fall into the five following groups: (1) the group comprising 32 organisms with the elongation efficiency determined only by the codon composition; (2) the group of 8 organisms with the codon composition inessential for the elongation efficiency; on the contrary, the elongation is limited by the degree of local secondary structures; however, the energies of these local secondary structures play no role; (3) the group consisting of only one organism—*P. aeruginosa* PA01—with no effect of codon composition on elongation efficiency, controlled by both the local secondary structures and their energies; (4) the group of 30 organisms where both the codon composition and local secondary structures are essential for determining the elongation efficiency; however, the energies of the local structures are insignificant; and (5) the group of 3 organisms where all the three factors—codon composition, local secondary structure, and their energies—are essential.

Acknowledgments

The work was supported in part by the Russian Foundation for Basic Research (grants № 02-04-488802, 01-07-90376, and 02-07-90359); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Likhoshvai V.A., Matushkin Yu.G. (2002). Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy. FEBS. 516:87-92.
2. Turner D.H., Sugimoto N. (1988). RNA structure prediction. Ann. Rev. Biophys. Biophys. Chem. 17:167-192.

STUDY OF THE SPECIFIC CONTEXTUAL FEATURES OF TRANSLATION INITIATION AND TERMINATION REGIONS IN EUKARYOTES

[1] Vishnevsky O.V., [2] Avdeeva I.V., [1] Kolchanov N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[2] Novosibirsk State University, Novosibirsk, 630090, Russia
[1] Corresponding author
e-mail: oleg@bionet.nsc.ru

Key words: *structure–function organization of regulatory regions of eukaryotic mRNA, weight matrices*

Summary

Motivation: Understanding of the structure–function organization of 5'- and 3'-untranslated mRNA regions and their involvement in translation regulation in eukaryotes is required for reconstruction of the general pattern of the expression of genetic information in the cell, prediction of mRNA expression patterns, and solution of biotechnological and gene engineering problems.

Results: We have performed a computer analysis of sequences near the initiation and termination codons of high- and low-expressed mRNA fractions of *Saccharomyces cerevisiae* from the TransTerm database (Dalphin et al., 1997). These regions were compared in high- and low-expressed mRNA fractions and described in terms of trinucleotide weight matrices. A correlation between the context organizations of the 5'- and 3'-untranslated regions of high-expressed genes has been revealed. Computer evolution simulation with the use of genetic algorithm has shown that this correlation can be explained by the limiting link model.

Introduction

Posttranscriptional control is an important link in the regulation of expression of genetic information. The rate of mRNA translation is greatly affected by the structure of the 5'- and 3'-untranslated regions, contextual features, and the presence of specific regulatory elements (Ray et al., 1983). The 5'-untranslated regions of many viral and some cellular mRNAs contain translation enhancers, which increase significantly the rate of mRNA translation (Gallie et al., 1996). Regulatory signals located in 3'-regions of eukaryotic mRNAs affect the stability, intracellular localization, and efficiency of translation initiation of mRNA (Wang et al., 1995).

The goal of this study is the analysis of context features of sequences near the initiation and termination codons of high- and low-expressed *Saccharomyces cerevisiae* mRNA fractions retrieved from the TransTerm database (Dalphin et al., 1997). We have constructed trinucleotide weight matrices describing the context of 5'- and 3'-untranslated mRNA regions. On their grounds, a correlation between the context organizations of the 5'- and 3'-regions has been discovered. Computer evolution simulation demonstrates that this correlation can be explained by the limiting link model.

Methods and Algorithms

We investigated sequences of the regions of the initiation and termination codons for 6741 mRNAs of *Saccharomyces cerevisiae* from the TransTerm database (Dalphin et al., 1997). Sequences 30 bp in length, from –21 to +9 with reference to the transcription start, and 30 bp from –9 to +21 with reference to the termination codon were examined. For each mRNA, the codon adaptation index (CAI) was retrieved from the same database. This index is known to reflect the correspondence between the frequency distribution of synonymous codons in mRNA coding sequences and the concentration of major tRNA fractions in the cell (Sharp et al., 1987). For *Saccharomyces cerevisiae*, CAI is a promising marker of mRNA expression in the cell. Four samples of 245 sequences each were constructed on the basis of CAI values for the 5'- and 3'-regions of high-expressed (CAI>0.5) and low-expressed (CAI<0.055) mRNAs.

The positional context of the regulatory regions of the mRNAs was estimated with the use of trinucleotide weight matrices. Positional weights were calculated as:

$$W(b,k) = \log[P_p(b,k)] - \log[P_n(b,k)],\ b \in A,\ k=1..L, \quad (1)$$

where $P_p(b,k)$ is the frequency of occurrence of trinucleotide at the position k of a training sequence sample; P_n, frequency of occurrence of this trinucleotide at this position in a sample of negative sequences; A, set of all trinucleotide values; and L, length of the weight matrix. The value $W(b,k)$ was assumed to be equal to –2 if $W(b,k)<-2$, and 2, if $W(b,k)>2$. Two types of weight matrices were constructed: (1) $M_{rand\_3'}$ and $M_{rand\_5'}$, for which a sample of randomly generated sequences

with neutral mononucleotide composition was used as negative, and (2) $M_{contr\_3'-}$ and $M_{contr\_5'-}$, for which samples of sequences contrasting in expression level were used as negative.

The score of an unknown sequence S of the length L during recognition by a weight matrix W was calculated as sum of the weights of corresponding positions:

$$S_L = \Sigma_{k=1...L} W(b_k, k). \quad (2)$$

Results

Trinucleotide weight matrices were constructed for description of the transcription initiation ($M_{rand\_5'-}$) and termination ($M_{rand\_3'-}$) in high- and low-expressed mRNA fractions in comparison with random sequences. It was found that some trinucleotides were absent from the 5'-regions of high-expressed mRNAs. In particular, ATG is lacking in the nearly whole 5'-region of such mRNAs. On the contrary, this exclusion is virtually absent from low-expressed mRNAs. For example, the AUG codon is entirely absent from position –11 with reference to the initiation codon ($W_{(AUG,-11)} = -2$), whereas in low-expressed mRNAs the frequency of this codon in this position slightly exceeds the random value ($W_{(AUG,-11)}$=0.1).

This confirms the commonly known fact that the presence of multiple ATGs near the transcription start can bring about false translation starts, which is extremely unfavorable for high-expressed mRNAs (Kochetov et al., 1998). Some differences in positional weights also occur in the context of the termination codon. For example, trinucleotides TAA and TAG are entirely absent from the –8 region with reference to the termination codons of high-expressed mRNAs ($W_{(TAA,-8)} = -2$, $W_{(TAG,-8)} = -2$), whereas this exclusion was not observed in low-expressed mRNAs ($W_{(TAA,-8)} = 0.04$, $W_{(TAG,-8)} = 0.13$).

Table. Positional nucleotide weights for the regions of (**a**) initiation translation context $M_{rand\_5'-}$ and (**b**) termination codon context $M_{rand\_3'-}$ in high- and low-expressed mRNA fractions. Positions with the complete absence of the corresponding trinucleotide are boldfaced.

a)

Position of start	ATG (high)	ATG (low)
–21	–2	–2
–20	–2	–0.38
–19	–2	0.1
–18	–2	–2
–17	–2	–0.22
–16	–2	–0.7
–15	–2	–0.2
–14	–2	–0.38
–13	–0.53	0.1
–12	–2	–2
–11	–2	0.13
–10	–0.5	–0.12
–9	–2	–0.69
–8	–2	–0.1
–7	–2	–0.32
–6	–2	–0.71
–5	–2	–0.68
–4	–2	–0.12
–3	–2	–2
–2	–2	–2
–1	–2	–2
0	1.87	1.66
1	–2	–2
2	–2	–2
3	–2	0.14
4	–2	0.08
5	–2	–0.39
6	–0.04	0.26

b)

Position of stop	TAA (high)	TAG (high)	TGA (high)	TAA (low)	TAG (low)	TGA (low)
–8	–2	–2	0.32	0.04	0.13	0.16
–7	0.28	–0.62	0.11	–0.1	–0.08	–0.13
–6	–2	–2	–2	–2	–2	–2
–5	–0.81	–2	0.15	0.09	0.04	0.04
–4	–0.45	–0.37	0.31	–0.05	–0.2	0.04
–3	–2	–2	–2	–2	–2	–2
–2	–2	–2	–2	–2	–2	–2
–1	–2	–2	–2	–2	–2	–2
0	1.23	0.83	0.47	1.1	1.13	1.24
1	–2	–2	–2	–2	–2	–2
2	–2	–2	–2	–2	–2	–2
3	–1.06	–2	–0.37	0	–0.43	0.21
4	–0.1	0.02	0.12	0.14	–2	0.04
5	–0.07	0.01	0.01	0.01	–2	–0.12
6	0.11	–0.1	–0.06	0.1	–0.04	–0.17
7	–0.59	–2	0.37	0.12	–0.71	0.06
8	–0.14	–0.38	–0.11	0.19	–0.13	–0.41
9	0.46	0.15	0.63	–0.31	0.11	–0.17
10	–0.47	–2	0.07	–0.27	–0.21	–0.38
11	–0.1	–0.24	0.09	0.07	–0.38	–0.67
12	0.01	–0.22	–0.01	0.03	–0.09	0.13
13	–0.24	–0.67	0.03	0.04	–0.69	–0.71
14	–0.06	0.1	0.02	–0.32	0.03	–0.14
15	0.21	–0.67	0.35	0.02	–0.09	–0.01
16	–0.32	–0.7	–0.34	–0.13	0.25	–0.23
17	0.01	–2	–0.02	–0.2	–0.04	–0.39
18	–0.02	–0.16	–0.13	–0.88	–0.08	0.14
19	–0.37	–0.07	0.02	–0.25	–0.08	–2

To reveal all positional differences between 5'- and 3'-untranslated mRNA regions with contrasting mRNA expression levels, we have constructed contrasting weight matrices $M_{contr\_5'}$ and $M_{contr\_3'}$. Their analysis shows significant positional differences in the regulatory regions of high- and low-expressed mRNAs. The dependence of CAI of yeast mRNA on the context of (a) 5'- and (b) 3'-untranslated mRNA regions is shown in Fig. 1. Note that no relation between these values is observed when the total set of mRNA sequences is considered. However, the fraction of high-expressed mRNAs (CAI>0.3) shows significant correlations between the context of the 5'- and 3'-regions and CAI. This threshold (CAI>0.3) was reported by Kochetov et al. (2001), who showed that the oligonucleotide composition of the 5'-regions of yeast mRNAs with CAI ~ 0.3 behaves in a similar manner, in comparison with all other sequences. This means that the value CAI>0.3 separates the high-expressed yeast mRNA fraction from mRNAs of all the other fractions. Thus, in high-expressed mRNAs, the context of the translation start and mRNA coding sequence correlate with the context of the termination codon and mRNA coding sequence.

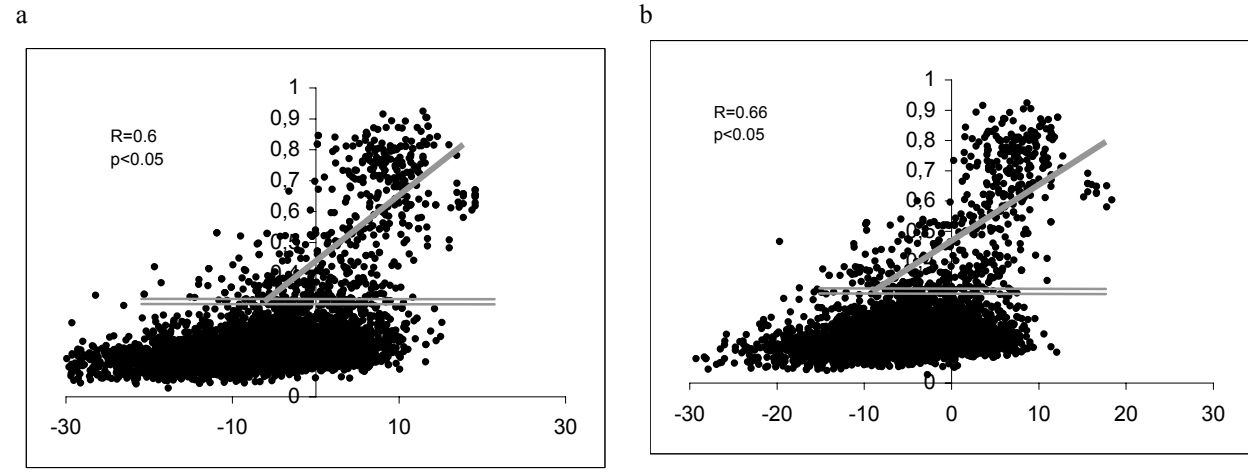


Fig. 1. Dependence of CAI (Y axis) on the context of (**a**) the AUG codon and (**b**) the termination codon (X axis). High-expressed mRNAs (CAI>0.3) are separated from the other mRNA fractions with a double horizontal line. Linear regression dependencies have been constructed for these mRNAs.

Of special interest is the correlation between the context of the initiation and termination codons. Figure 2 (filled circles) shows that it is significant. Thus, the evolution of yeast genomes established correlations between the contexts of the initiation codon, coding region, and termination codon for high-expressed mRNAs.

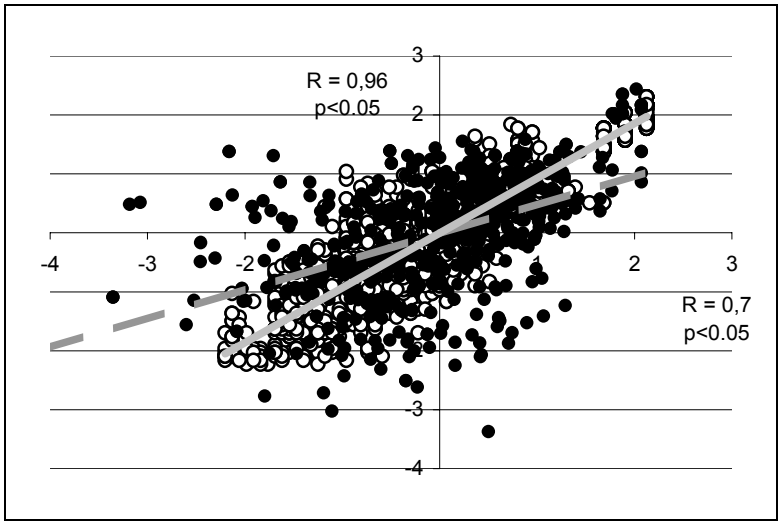


Fig. 2. Filled circles: correlations between the contexts of the AUG codon (X axis) and termination codon (Y axis) for the yeast mRNA fraction with CAI>0.3. Open circles: correlations the AUG codon and termination codon for a mRNA with CAI>0.3 simulated with the use of the genetic algorithm. White dash line: a regression line describing the behavior of yeast mRNAs. White solid line: a regression line describing the behavior of the computer-simulated mRNA.

We applied genetic algorithm as a simulation method for understanding the mechanisms of emergence of these correlations. The evolution of a population of mRNA sequences, including the 5'-region, coding region, and 3'-region was considered. It was governed by (1) recombinations, which exchanged fragments of 5'-, coding, and 3'-regions among mRNA molecules; (2) point mutations; and (3) selection directed to the increase in translation rate F according to the limiting link model.

The rate of mRNA translation F was assumed to be determined as:

$$F = \min\begin{cases} Score(5'\text{–}region) \\ CAI(coding\_region) \\ Score(3'\text{–}region) \end{cases}. \quad (3)$$

Here, *Score(5'-region)* and *Score(3'-region)* are scores of the 5'- and 3'-regions calculated from the corresponding weight matrices and *CAI(coding_region)* is the aforementioned codon adaptation index.

Translation of mRNA includes a succession of three processes: initiation, elongation, and termination. Obviously, the greater is each of the indices *Score(5'-region)*, *Score(3'-region)*, and *CAI(coding_region)*, the higher is the rate of each of the three processes. However, according to the limiting link model, the overall efficiency of mRNA translation is limited by the rate of the slowest process, that is, the least efficient link.

The simulation shows (Fig. 2, open circles) that at CAI>0.3, the behavior of natural mRNAs reflects the actual correlation between the context features of the 5'- and 3'-untranslated regions of high-expressed mRNAs.

Discussion

Posttranscriptional regulation, which determines the translation rate of a particular RNA, involves translation initiation, elongation, and termination. We have described it according to a model of an unbranched sequential molecular process involving three stages, which determine the yield of the final product, protein. The model of limiting link is a good approximation to such linear processes. Its application to biological systems and processes was most comprehensively described by Poletaev (1973) and Ratner (1990).

Limiting links are those links in a chain of reactions, which determine the yield of the final product of the whole chain. Thus, the change in the reaction rate in the limiting link can alter significantly the yield of the final product, whereas changes in nonlimiting links do not bring about notable yield changes. Hence, control of the limiting link of a system is the most efficient method for controlling the whole system (Ratner, 1990). As soon as a mutation makes a certain link nonlimiting, the regulation is determined by the next limiting link. Description of evolution of mRNA molecules according to this model means that the overall efficiency of the translation system can be increased only by mutations removing such restrictions. Other (nonlimiting) elements of mRNA context organization undergo neutral evolution independently of one another.

Our results show that the evolution of high-expressed mRNAs optimized three main mRNA regions affecting translation rate: the 5'-untranslated, coding, and 3'-untranslated regions. This explains the correlation of context features of these regions in high-expressed mRNAs (CAI>0.3), which have experienced the greatest selection pressure directed to the increase in translation rate. We observe this by the example of yeast (Fig. 2, filled circles). It is the sample of high-expressed mRNAs with CAI>0.3 that exhibits a significant correlation between the scores of the 5'- and 3'-untranslated mRNA regions (Fig. 2)

A low translation level can be determined by the presence of a limiting link in any of the three regions. In this case, mRNAs with good context properties of the translation start and coding region may have a low overall translation rate because of the presence of a limiting link in the termination codon. Similarly, the presence of a limiting link at the level of a coding codon may determine a low translation rate even with good contexts of the initiation and termination codons and so on. Obviously, in this case of low-expressed mRNAs, correlation of context properties of the three regions is unlikely.

Some differences between the actual and theoretical correlations can be presumably explained by incomplete consideration of fine specific features in the dependence of the overall translation rate on the relations between contributions of the three components according to the limiting link model in functional (3). Note that this approach can also be used in the inverse problem: evaluation of the selection pressure on the evolution of the contexts of 5'-, coding, and 3'-untranslated mRNA regions.

Acknowledgements

The authors are grateful to A.V.Kochetov and V.A.Likhoshvai for supplied materials and fruitful discussions. The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 07-90337, 00-02-07-90355, 00-04-49229, and 00-04-49255); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); US National Institute of Health (grant № 2 R01-HG-01539-04A2): and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Dalphin M.E, Brown C.M., Stockwell P.A., Tate W.P. (1997) Nucl. Acids Res. 25, 246–247.
2. Gallie D.R. (1996) Plant Mol. Biol. 32, 145–158.

3. Kochetov A.V., Ischenko I.V., Vorobiev D.G., Kel A.E., Babenko V.N., Kisselev L.L., Kolchanov N.A. (1998) FEBS Lett. 3, 351-355.
4. Kochetov A.B., Grigorovich D.A., Titov I.I., Vorobiev D.G., Syrnic O.A., Vishnevsky O.V. (2001) Mol.r Biol. (Mosc.). 6, 1039-47.
5. Poletaev I.A. (1973) Zurnal Obshej Biologii (Rus.). 34, 43.
6. Ratner V.A. (1990) Genetika (Rus.). 5, 789-803.
7. Ray B.K., Brandler T.G., Adya S., Daniels-McQeen S., Miller J.K., Hershey J.W.B., Grifo J.A., Merrick W.C., Thach R.E. (1983) Proc. Natl Acad. Sci. USA. 80, 663–667.
8. Sharp P.M, Li W-H. (1987) Nucl. Acids Res. 15, 1281–1295.
9. Wang S., Miller W.A. (1995) J. Biol. Chem. 270, 13446.

THEORETICAL ANALYSIS OF TRANSLATIONAL EFFICIENCY OF THE AQUAPORIN 4 mRNA ISOFORMS

[1] Alikina T.Y., [1,2] Zelenin S.M., [1] Bondar A.A.*

[1] Novosibirsk Institute of Bioorganic Chemistry, Novosibirsk, 630090, Russia, e-mail: alikina@niboch.nsc.ru
[2] Karolinska Institutet, Q2:09 ALB KS, S-17176 Stockholm, Sweden

***Key words**: aquaporin 4 (AQP4, MIWC), 5'-UTR, mRNA, translational efficiency, water channel*

Resume

Motivation: The water channel aquaporin 4 (AQP4) is expressed in brain, kidney, lung, and muscle and has been suggested to play an important role in the regulation of water homeostasis. There are at least three different mRNAs (M1, M23X and M23) encoding two isoforms of mouse AQP4 (M1 and M23) known by now (Ma et al., 1996; Turtzo et al., 1997; Zelenin et al., 2000). These mRNAs differ in expression pattern (Zelenin et al., 2000). We suppose that the differences in sequence feautures of AQP4 mRNA's 5'-UTRs correlate with the differences in their expression patterns and, finally, with the gene expression regulation by different promoter regions.

Results: Here we theoretically estimated a translation properties of identified by us AQP4.M1 mRNA (AF469168) and AQP4.M23X mRNA (AF469169), and previously described MIWC1 (M23) (Ma et al., 1996) mRNA encoding mouse aquaporin 4 water channel by use the computer expert system Leader_RNA (http://wwwmgs.bionet.nsc.ru/mgs/gnw/leader/). We suppose that under specific conditions or in certain tissues in addition to AQP4 mRNA having low translation efficiency cells could express AQP4 mRNA possessing much higher translation efficiency. We found that AQP4 mRNA forms differ in predicted translational activity which should be taken into account for further characterization of the AQP4 expression pattern.

Introduction

Aquaporin 4 (AQP4), a transmembrane protein – water channel, has been suggested to play an important role in the regulation of water homeostasis (Ma et al., 1996; Turtzo et al., 1997). In kidney AQP4 present in basolateral membrane of the collecting duct principal cells and play, together with AQP2 and AQP3, an important role in the regulation of water reabsorption (Nielsen et al., 2002). Until recently AQP4 has been suggested to be a single water channel protein that expressed in brain. Now it is demonstrated that in the central nervous system AQP4, 3, 5, 8 and 9 are expressed (Yamamoto et al., 2001). AQP4 is supposed to mediate transmembrane water movement at the blood-brain barrier and brain-cerebrospinal fluid interface, take part in osmosensor processes and seems to be a major molecule in the cause of brain edema. Mechanisms of AQP4 gene expression and regulation of protein water permeability has been a little studied (Han et al., 1998; Nakahama et al., 1999; Zelenin et al., 2001; Yamamoto et al., 2001b). It was demonstrated recently that expression of AQP4 M1 and M23 mRNAs, and synthesis of AQP4 protein is regulated in cultured astrocytes during hypoxia and reoxygenation (Yamamoto et al., 2001b). Also it was demonstrated that TPA (PKC activator) decreases the expression of AQP4 mRNA at the transcriptional level in astrocytes (Yamamoto et al., 2001a; Nakahama et al., 1999). Earlier it has been considered that there were at least two forms of AQP4 mRNA (M1 and M23) that encode two isoforms of the AQP4 protein, M1 and M23 respectively. We revealed the presence of a third mouse AQP4.M23X (AF469169) mRNA and demonstrated that two of three forms of AQP4 mRNA (M23 and the new one M23X) have a tissue- and age-specific expression (Zelenin et al., 2000). New AQP4.M23X mRNA has the start of transcription in the new exon X of mouse AQP4 gene and encodes previously described AQP4 isoform, M23 (Bondar et al., 2000). M1 isoform only contains 22 more amino acids in the N-terminal than M23 (Neely et al., 1999), which have been shown do not influence on water permeability of the channel. What could be a key to the explanation of why different mRNAs (M23 and M23X) encoding the same AQP4 protein or the protein isoform (M1) exist and to their probable functional role? We suppose that it could be differences in translational efficiency of the AQP4 mRNA isoforms characterizing with different tissue- and age-specific expression pattern (Zelenin et al., 2000).

Prediction of the gene expression pattern through computational analysis of the nucleotide sequence is still one of the main tasks of modern bioinformatics. Accurate prediction is very complicated because the level of eukaryotic gene expression may be regulated at various steps: transcription, pre-mRNA processing and export, mRNA translation, and the cytoplasmic stability of the mRNA and polypeptide (Kochetov et al., 1999). Contextual and structural features of the gene nucleotide sequence may influence the efficiency of expression at all stages, therefore they should be all considered in detail. Analysis of mRNAs translatability in eukaryotic cells is one of the particular tasks in the framework of this general problem (Kochetov et al., 1999). It is well known that translational efficiency of eukaryotic mRNAs varies considerably with their

* Corresponding author.

sequence characteristics. 5'UTR features such as an ability to form secondary structures, G+C content, false start codons within 5'UTR and others may have influence on the efficiency of mRNA translation. It is widely accepted that the majority of eukaryotic mRNAs are translated through the linear scanning mechanism (Kozak, 1994). According to this model, several features of the leader sequence influence mRNA translational efficiency, i.e. the context of the translational start codon, occurrence of AUGs within 5'UTR, and the stable secondary structure in the leader. Recently such computer tool for the prediction of mRNA translatability taking into account 5'UTR features of mRNA have been developed at the Institute of Cytology and Genetics of SD RAS (Kolchanov et al., 2001).

Methods and Algorithms

We described cloning of mouse AQP4 gene and it's new exon prediction and experimental confirmation earlier (Zelenin et al., 2000; Bondar et al., 2002). Translation properties of mRNAs were estimated by using ICG SB RAS server expert system {{http://wwwmgs.bionet.nsc.ru/mgs/gnw/leader/}} (Kolchanov et al., 2001).

Results and Discussion

It is possible that the existence of three AQP4 mRNA isoforms and the different pattern of their age- and tissue-specific expression (Zelenin et al., 2000) are connected with differences of their stability in the cells or their translation efficiency. Here translation properties of identified by us AQP4.M1 mRNA (AF469168) and AQP4.M23X mRNA (AF469169), and previously described MIWC1 (M23) (Ma et al., 1996) mRNA were estimated theoretically by use of computer expert system {{http://wwwmgs.bionet.nsc.ru/mgs/gnw/leader/}} that allows to predict mammalian mRNA translation behavior by use of a detailed analysis of 5'UTR structure features (Kochetov et al., 1999; Kolchanov et al., 2001). It is well known that contextual and structural features of the 5'UTR have significant influence on the rate of translation initiation and, thereby, on the level of polypeptide production and in that way could influence on the water flux across the cell membrane via abundance of aquaporins. According computer prognosis mouse AQP4.M1 mRNA has low translation efficiency (coef.=-0.189765). On the other hand, MIWC1 (M23) mRNA which has a transcription initiation site in exon 1 located more than 3500 b.p downstream from exon 0 has high translation efficiency (coef.=+0.191242). It is still unclear what regulatory regions of the AQP4 gene control the expression of the third, AQP4.M23X, mRNA. It is unclear also if there are transcription factors binding sites between exons 0 and X which themselves regulate AQP4.M23X mRNA transcription. We suppose that probability that promoter 0, which is located at only about 700 b.p distance upstream exon X (for the detailed scheme of AQP4 mRNA formation see (Bondar et al., 2002)), have an essential influence on AQP4.M23X mRNA expression is high enough. Moreover it appears that AQP4.M23X mRNA, which has transcription initiation site in exon X, has also high translation efficiency (coef.=+0.182715). We suppose that under some conditions or tissue specifically together with AQP4 mRNA with low translation efficiency cells can express AQP4 mRNA possessing much higher translation efficiency. This may result in increase of AQP4 level and therefore in increase of water permeability of the cell membrane.

Recently we determined nucleotide sequence of 8629 b.p. fragment upstream of mouse AQP4 gene exon 0 and it appears that there could be an another one AQP4-like mRNA form transcribed from AQP4 gene. We suppose that AQP4-like mRNA may be transcribed from at least two new exones (A and B) . By comparative analysis of AQP4-like mRNA and mouse AQP4 gene we shown that this mRNA doesn't contain nucleotide sequence corresponding to exon 0, X, 2, 3 but it does contain putative exons A and B, parts of exons 1 and 4. Splicing site between exon B and exon 1 is identical with splicing site known for earlier identified mRNA M1 (AF469168) (between exons 0 and 1) и mRNA M23X (AF469169) (between exons X and 1). The functional role and expression pattern of the AQP4-like mRNA as well as a protein function it encodes are unknown. It may be proposed that this mRNA may encode small highly homologous to AQP4 protein or for example protein resembling "agglutinated" together AQP4 fragments. The AQP4-like mRNA may also represent an alternative splice variant of unknown AQP4 pre-mRNA common for all already described AQP4 mRNA forms.

A quite complicated structure-functional organization of the AQP4 gene assumes also that highly organized mechanisms regulating expression of each AQP4 mRNA isoform should exist. It could be supposed that this "complexity" reflects a presence of several parallel regulatory mechanisms when one could play a reserve role in a case of malfunction of other. It also could reflect the importance of AQP4 water channel for water homeostasis maintanance, since a failure of its functioning can result in serious consequences for organism. May be the differences in translational efficiency of AQP4 mRNAs that is obviously influenced by 5'-UTR structure could be a key to the explanation of differences in their expression pattern and finally to the gene expression regulation by different promoter regions.

This work was supported by the RFBR 01-04-49390 grant.

References

1. Bondar A.A., Alikina T.Y. et al. (2000) Mouse aquaporin 4 gene: prediction of a new exon and experimental confirmation. In SB RAS ICG (eds) Proc. of the Second Intern. Conf. on Bioinformatics of Genome Regulation and Structure. Novosibirsk, August 7-11, 2000. 2, 46-48.
2. Bondar A.A., Alikina T.Y. et al. (2002) Structure-functional organization of mouse aquaporin 4 gene. Izv. Acad. Nauk. Ser. Chim. In press.

3. Han Z., Wax M.B. et al. (1998) Regulation of aquaporin-4 water channels by phorbol ester-dependent protein phosphorylation. J. Biol. Chem. 273(11), 6001-6004.
4. Kochetov A.V., Ponomarenko M.P. et al. (1999) Prediction of eukaryotic mRNA translational properties. Bioinformatics. 15(7-8), 704-712.
5. Kolchanov N.A., Titov I.I. et al. (2001) Computational approaches to RNA structure and function analysis. In SB RAS NIBC (eds) Abstracts of Intern. Conf. "RNA as Therapeutic and Genomic Target". Novosibirsk, 30th August-2nd September 2001. 38.
6. Kozak M. (1994) Determinants of translational fidelity and efficiency in vertebrate mRNAs. Biochimie. 76, 815-821.
7. Kozak M. (1996) Interpreting cDNA sequences: some insights from studies on translation. Mamm. Genome. 7(8), 563-574.
8. Ma T., Yang B. et al. (1996) Gene structure, cDNA cloning, and expression of a mouse mercurial-insensitive water channel. Genomics. 33, 382-388.
9. Nakahama K., Nagano M. et al. (1999) Effect of TPA on aquaporin 4 mRNA expression in cultured rat astrocytes. Glia. 25(3), 240-246.
10. Neely J.D., Christensen B.M. et al. (1999) Heterotetrameric composition of aquaporin-4 water channels. Biochemistry. 38, 11156-11163.
11. Nielsen S., Frokier J. et al. (2002) Aquaporins in the Kidney: From Molecules to Medicine. Physiol. Rev. 82, 205–244.
12. Turtzo L.C., Lee M.D. et al. (1997) Cloning and chromosomal localization of mouse aquaporin 4: exclusion of a candidate mutant phenotype, ataxia. Genomics. 41, 267-270.
13. Yamamoto N., Sobue K. et al. (2001a) Differential regulation of aquaporin expression in astrocytes by protein kinase C. Brain Res. Mol. Brain Res. 95(1-2), 110-116.
14. Yamamoto N., Yoneda K. et al. (2001b) Alterations in the expression of the AQP family in cultured rat astrocytes during hypoxia and reoxygenation. Brain Res. Mol. Brain Res. 90(1), 26-38.
15. Zelenin S., Gunnarson E. et al. (2000) Identification of a new form of AQP4 mRNA that is developmentally expressed in brain. Pediatr. Res. 48(3), 335-339.
16. Zelenina M., Zelenin S. et al. (2002) Water permeability of aquaporin 4 is decreased by protein kinase C and dopamine. Amer. J. Physiol. Renal. Physiol. Articles (in press).

A COMPUTER DIFFERENTIAL DISPLAY REVEALS GENES WITH SPECIFIC EXPRESSION PATTERNS: FROM POTENTIAL HUMAN TUMOR MARKERS DOWN TO PLANT STRESS-RESISTANCE GENES

[1] Baranova A.V., [1] Lobashev A.V., [1] Ivanov D.V., [2] Krukovskaya L.L., [3] Zinchenko V.V., [3] Shestakov S.V., [2] Kozlov A.P., [1] Yankovsky N.K.

[1] Vavilov Institute of General Genetics, Moscow, Russia, e-mail: baranova@vigg.ru
[2] Biomedical Center, Research Institute of Pure Biochemicals, St.-Petersburg, Russia
[3] Department of Genetics, Moscow State University, Moscow, Russia

Key words: *differential display, human tumor marker, plant stress marker, computer analysis*

Resume

Motivation: One of the most important tasks of the current molecular biology is an investigation of gene expression patterns in the different kinds of cells. The state of molecular-genetic databases requires the efforts directed towards the high throughput computer analysis rather than a blind accumulation of experimental data. Thus it is necessary to develop effective computational approaches to the analysis of gene expression data i.e. various implementations of Differential Display.

Results: We have developed a software called HSAnalyst and searched a publicly available NCBI database (UniGene) to find 197 human genes that are presented in the database by ESTs preferntially (>10 times) from tumor sources. The expression of the genes can possibly serve as potential human tumor development marker. The software was also successfully applied to find three of A.thaliana genes that are differentially expressed under stress conditions.

Introduction

The comparison of gene expression patterns in different cell or tissue types is a powerful and reliable method of the modern molecular biology. One of the most important goals is a search for human genes that have increased expression level in tumor cells compared to normal ones. There is a number of experimental approaches to solve such a problem as a search for tumor-specific genes. The approaches are rather expensive and require a lot of resources (for example – modifications of differential display, cDNA microarray analysis or SAGE – serial analysis of gene expression) (Green et al., 2001; Carulli et al., 1998).

A computer-based procedure can be used to compare different expression profiles. Such kind of method is often referred as computer differential display (CDD). The CDD principle of comparing the gene expression profiles is the following. A level of a gene expression in a particular condition (or tissue) is presented by the number of EST found in electronic dbEST in the cluster of different ESTs corresponding to the gene.

The number of human ESTs in publicly available databases (>$3x10^6$) exceeds by approximately two orders of magnitude the number of known human genes ($2,5–4x10^4$) (Craig Venter et al., 2001; Lander et al., 2001). The ratio allowed a successful implementation of the CDD method (Vasmatzis et al., 1998; Scheurle et al., 2000) to find genes differentially expressed in some human tissues compared to the corresponding normal tissues.

Implementation and Results

We have tried to implement the CDD approach to search for human genes that would mark a tumour by its expression. This gene-hunting procedure was inspired by the hypothesis that tumours may provide conditions for the expression of some transcribed units that are not expressed in any normal tissues (Kozlov, 1996). We performed a differential displaying of ESTs from a pool of all available tumour libraries against a pool of all available normal libraries instead of pairwise comparison of each tumour and corresponding normal tissues (Evtushenko et al., 1989).

There is free available online software on the NCBI servers implementing CDD algorithms (UniGene and CGAP departments). The main disadvantage of this software is a limited access to the raw data. To overcome this limitation we had to use the full set of source data also available on the NCBI server arranged in a few plain-text files. These data describe the clusterisation of all human ESTs in approximately 90000 sets (clusters) which presumably correspond to different genes. ESTs in these clusters are grouped according to shared stretches of nucleotide sequences. By the time of this work the database consisted of $2,2x10^6$ descriptions of individual cDNA clones together with source tissues references. We have developed a program called HSAnalyst to classify data from original dbEST and UNIGENE databases. We have found a number of clusters that were formed exclusively or overwhelmingly by ESTs from tumor tissues but not from the normal ones. Products of such genes could possibly serve as new highly informative tumor markers.

UniGene database build 129 was used for EST sorting. On the first hand we were interested in each of approximately 90000 EST clusters and in the information about all the clone libraries we could access via the web-interface from UniGene. We have received the most part of the needed data from NCBI server ftp://ncbi.nlm.nih.gov/repository/UniGene/ via FTP protocol. We have developed a program called HSAnalyst to classify data from original dbEST and UNIGENE databases in a table form. This program uses our own created clone libraries descriptions database that we have built after thorough and detailed analysis of the available human clone libraries descriptions.

We have checked the description of each clone library using the resources of dbEST, UniGene, CGAP databases located on NIH, TIGR and Stratagene web-servers and some of information was retrieved from the NCBI electronic article database PubMed. Every clone library descriptions in our database were classified as "tumor" and "normal" according to its tissue source or the source cell culture type (derived from normal or tumor tissue). Also there were "unclassified" libraries which appeared to have no reliable description or were prepared using inadequate experimental approaches like e.g. the microdissection method in which it is hard to make sure that there are no "normal" cells in a bulk of a tumor tissue sample taken from the edge of the tumor body. After all we have identified 2681 clone libraries as "tumor", 1087 libraries as "normal" and 227 as "unclassified".

An algorithm executed by the program consists of two major steps: 1) for each cluster the number of its ESTs is retrieved from cluster description and 2) the number of ESTs from the "tumour" cDNA libraries is counted according to the LibraryRegistry database. The whole range of possible EST numbers is dissected into subranges. HSAnalyst makes possible to arrange subranges exponentially (subranges with exponents 1-2, 3-4, 5-8, 9-16, etc) or linearly (subranges with factors 1-10, 11-20, 21-30 etc). Simultaneously the ratio cancer ESTs/all ESTs is calculated for each cluster and those, which exceed the user-defined bottom threshold value, are listed in the output file. To be sure that we have found a "true" tumour-specific clusters not generated by chance among the great total number of the EST clusters (more than 90000 units) we have calculated the theoretical number of "tumour" clusters for every subrange. The underlying model is the binomial distribution with the mean value of "cancer/all" ratio that can be declared by user (0 to 100%) The number of clusters that exceed threshold value is calculated. For each range of cluster sizes, "tumor-related" ESTs content, the number of clusters that comply with the input conditions and the expected number of such clusters are calculated.

For the current database, the mean content of "cancer-related" ESTs is about 48%. The most interesting were tumor-related clusters with size of 16-128 ESTs and satisfying the pre-defined conditions (more than 10 ESTs in cluster and the tumor/total ratio is greater than 90%). Twenty-one clusters falling in range of 16-128 contains only tumour-derived ESTs and fulfills a threshold 100%.

Some of the found clusters have a significant homology to known proteins. For example CDD revealed the HAND1 gene coding homeobox protein 1 from H6 family, OCIM homologue (90% of 344 a) involved in multiple myeloma development, mouse GDF3 homologue (84% of 96 a) and a gene similar to IQGAP1 that coding a protein homologous to RAS GTPase activating protein. Possible involvement of these genes in human tumor development requires further investigation.

The most interesting are clusters represented by ESTs found exclusively in the tumour-derived libraries. The striking feature of the analyzed tumour-specific clusters is their frequent occurrence in libraries from colon carcinomas (Hs. 560, Hs. 1085, Hs. 239891) or lung and ovarian carcinomas (Hs.145340, Hs. 145509, Hs. 181624, Hs. 293429, Hs. 133107, Hs. 133296, Hs. 145492, Hs. 181624). Interestingly, all three colon-specific EST clusters obtained by our analysis represent known genes encoding apolipoprotein B mRNA-editing protein APOBEC1, guanylate cyclase 2C and G protein-coupled receptor 35. Both APOBEC1 and guanylate cyclase 2C mRNAs have been shown to be overexpressed in colon carcinomas (Lee et al., 1998; Carrithers et al., 1996). Moreover, the high-level expression of APOBEC1 in transgenic mice and rabbit liver causes liver dysplasia and hepatocellular carcinomas (Yamanaka et al., 1995). mRNA encoding guanylate cyclase 2C appears to be a relatively specific marker of the presence of metastatic colonic carcinoma cells in normal tissues including peripheral blood (Carrithers et al., 1996). In our opinion, the gene encoding G protein-coupled receptor 35 deserves attention as a putative marker of colon cancer possibly involved in the progression of the disease.

EST clusters from lung and ovarian carcinoma libraries may also represent potential tumour markers. As far as they do not contain any homologies to known proteins and easily recognized open reading frames they may be considered as an evidence in favor of the expression of newly evolved DNA sequences in tumor cells (Kozlov, 1996) or as a manifestation of the phenomenon of the "background" or "illegitimate" gene expression (Chelly et al., 1989; Ko et al., 2000), which may be enhanced in tumor cells due to deregulation of the house-keeping processes.

The other task to the HSAnalyst was the CDD of a model plant Arabidopsis thaliana under stress conditions to find the genes differentially expressing under stress compared to a normal physiological state. The conducted analysis revealed statistically reliable quantitative differences between the gene expression in the normal plant and salt (hyperosmotic) stressed plant. We found 5 genes satisfying the following conditions: cluster has to consist of more than 10 ESTs and 80% out of all sequences must be expressed specifically under stress conditions. The most interesting of the found clusters are At.11290 corresponding to the glutation-S-transferase (GST30), At.5388 for the Lti30 and At.20845 corresponding to the cor15 polypeptide. The rest two of the found clusters correspond to the genes which functions are not established well and requires further investigation.

Discussion

Differentially expressed EST clusters (genes) may be useful as tumour markers and prognostic indicators and may be suitable targets for various therapeutic interventions. To meet the goal we have probed a subset of EST clusters by both confirmatory PCR and Northern experiments on Clontech Multiple Tissue cDNA Panels and MTN Northern blots. The results are reassuring as mRNA corresponding to one of the probed clusters, Hs. 133294, shows an expression in four different tumour cell lines but in none of the 16 normal tissues included in MTN blot. Cellular mRNAs that correspond to the most of the probed clusters are significantly overexpressed in at least two tumour samples in comparison to the normal tissues (data not shown), which confirm the validity of the CDD procedure applied in the article.

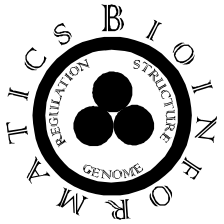
As for A.thaliana, our CDD method proved to be a reliable one for searching of stress-inducible genes that possibly could later be used for plant selection.

Acknowledgements

The work was supported by a grant from the state programm "Integration".

References

1. Green C.D., Simons J.F., Taillon B.E., Lewin D.A. (2001) J. Immunol Methods. 250(1-2), 67-79.
2. Carulli J.P., Artinger M., Swain P.M., Root C.D., Chee L., Tulig C., Guerin J., Osborne M., Stein G., Lian J., Lomedico P.T. (1998) J. Cell Biochem. Suppl. 30-31, 286-96.
3. Craig Venter J. et al. (2001) Science. 291, 1304-51.
4. Lander E.S. et al. (2001) Nature. 409, 860-921.
5. Vasmatzis G., Essand M., Brinkmann U., Lee B., Pastan I. (1998) Proc. Natl Acad. Sci. USA. 95(1), 300-4.
6. Scheurle D., DeYoung M.P., Binninger D.M., Page H., Jahanzeb M., Narayanan R. (2000) Cancer Res. 60(15), 4037-43.
7. Kozlov A.P. (1996) Medical Hypotheses. 46, 81-84.
8. Evtushenko V.I., Khanson K.P., Barabitskaya O.V., Emel'anov A.V., Reshetnikov V.L., Kozlov A.P. (1989) Mol. Biol. 23(3), 510-520.
9. Lee R.M., Hirano K., Anant S., Baunoch D., Davidson N.O. (1998) Gastroenterology. 115(5), 1096-103.
10. Carrithers S.L., Barber M.T., Biswas S., Parkinson S.J., Park P.K., Goldstein S.D., Waldman S.A. (1996) Proc. Natl Acad. Sci. USA. 93(25), 14827-32.
11. Yamanaka S., Balestra M.E., Ferrell L.D., Fan J., Arnold K.S., Taylor S., Taylor J.M., Innerarity T.L (1995) Proc. Natl Acad. Sci. USA. 92(18), 8483-7.
12. Chelly J., Concordet J.P., Kaplan J.C., Kahn A. (1989) Proc. Natl Acad. Sci. USA. 86(8), 2617-21.
13. Ko Y., Grunewald E., Totzke G., Klinz M., Fronhoffs S., Gouni-Berthold I., Sachinidis A., Vetter H. (2000) Oncology. 59(1), 81-8.



COMPUTATIONAL PROTEOMICS

DEVELOPMENT OF A STRATEGY FOR COMPUTER-ASSISTED SEARCHING FOR FUNCTIONALLY SIMILAR PROTEINS IN EVOLUTIONARILY DISTANT ORGANISMS

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

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

Key words: *databases, knowledge bases, computer analysis, functional proteomics, virtual cell*

Abstract

Motivation: Synaptonemal complex (SC), an universal ultrastructure that ensures the successful pairing and recombination of homologous chromosomes during meiosis in evolutionarily distant organisms, is build of non-homologous proteins. We aimed on developing a method of searching databases for genes that code for such non-homologous but functionally analogous proteins.

Results: Advantage was taken of the ultrastructural parameters of SC and the conformation of SC proteins responsible for these. Using data from literature, we found a highly significant correlation ($r=0.97$; $P< 0.001$) between the width of the SC central space and the length of the alpha-helix in the central domain of yeast normal and deleted Zip1p and mammalian SCP1, intermediate proteins that form transversal filaments in SC central space. Basing on this, we found the *Drosophila melanogaster CG17604* gene whose virtual protein meets the correlation requirement. Our finding has received experimental support in another lab. With the same strategy, we showed that the *Arabidopsis thaliana* and *Caenorhabditis elegans* genomes contain unique genes coding for proteins that also fit the above requirements.

Availability: Bogdanov et al., 2002a, b.

Introduction

Genome databases have accumulated and processed the data on the complete genome sequences of model eukaryotic organisms: yeast *S. cerevisiae,* nematode *C. elegans,* fruit fly *D. melanogaster*, and plant *A. thaliana*. These organisms have been found to possess several hundred of orthologous genes and proteins, which are similar in primary structure and play a common role. However, evolutionarily distant organisms have such organelles as kinetochores, cell centers, synaptonemal complexes (SCs), etc., which partly or completely differ in ultrastructure notwithstanding their common function. In many cases, these structures are build of different structural proteins. To search for such **functional analogs**, we developed a strategy that combines computer analysis of the conformation and other physical-chemical properties of proteins with ultrastructural parameters of cellular organelles *in situ* obtained by electron microscopy.

The transversal filaments (TFs) in the central space of the SC are responsible for chromosome synapsis (Zickler, Kleckner, 1999). Mammalian and yeast proteins that form TFs, SCP1 and Zip1p respectively, have been isolated and studied (Heyting, 1996; Dong, Roeder, 2000). These proteins, being non-homologous, are similarly organized and include three domains, with the central one possessing an extended alpha-helix. Both proteins are classed with intermediate proteins. *In vitro*, each protein forms road-shaped dimers of two similarly oriented parallel molecules (Heyting, 1996). The dimers resemble tooth-like halves of zipper-like connections in SC central space, i.e. TFs (Fig.). In addition, SCP1 and Zip1p share other physical-chemical properties of the entire molecule and of its individual domains. Their analogy can be extended to the ultrastructural level, since SC is structurally similar in yeast and in mammals, central space width, being about 100 nm. Basing on these data, we carried out a computer search for proteins forming TFs in *D. melanogaster, A. thaliana,* and *C. elegans*.

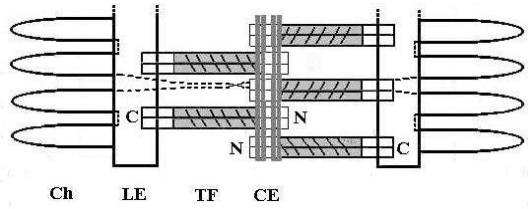


Fig. Scheme of synaptonemal complex in mammals .and yeast

Ch - Chromatin loops; LE - lateral element; TF - transversal filament; C and N -- terminal domains of SCP1 or Zip1p, respectively; CE - central element; doted line -- crossover DNA.

Method

As resources, we used databases on the known and putative genes and proteins in *S. cerevisiae, D. melanogaster, A. thaliana,* and *C. elegans*. provided by NCBI (http://www.ncbi.nlm.nih.gov/)..Additionally for *A. thaliana* and *C. elegans,* databases of **TAIR AGI Information** (http://www.arabidopsis.org/home.html) and **WormBase** (http://www.wormbase.org) were used respectively.

The analysis of protein domain structure and the search for structural/functional analogs was performed by the use of **CDART**:(Conserved Domain Architecture Retrieval Tool) (http://www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps). And it was extended by prediction of physical-chemical properties of proteins, using **ProtParam tool** provided by **ExPASy Molecular Biology Server** (Expert Protein Analysis System) available at (http://www.expasy.ch/tools/proprotparam.html), and prediction of the protein secondary structure (ISREC) provided by **BCM Search Launcher: Protein Secondary Structure Prediction** (http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html/).

Results and Discussion

Using the experimental data by Dong & Roeder (2000), we found that, in *zip1* mutants of *S.scerevisiae* having different deletions from the central domain of Zip1p, the TF length and central space width are well correlated with the length of the Zip1p normal either partially deleted alpha-helix (r=0.97, P<0.001). Along with certain protein features (the domain organization, the deduced conformation of the central domain, etc.), the correlation was used as a criterion to search for analogous proteins of *D. melanogaster* and other organisms. In fact, we sought the genes that potentially code for candidate SC TF proteins of these organisms.

The *c(3)G* mutation in *D. melanogaster* causes the same ultrastructural alterations in SC as *zip1* does in *S. cerevisiae*. Hence the virtual protein product of $c(3)G^{+}$ might be a good candidate for a *D. melanogaster* TF protein. We analyzed the virtual protein products of 78 *D. melanogaster* genes from the region covering the *c(3)G* locus (250 kb in section 88E-89B of chromosome 3R). The genes have been annotated by Celera Genomics Inc. (NCBI database). We found only one gene, *CG17604,* whose virtual protein product was similar to Zip1p and SCP1 by all the criteria used. The length of its alpha-helical region proved to correspond to the central space width in *D. melanogaster*. We identified the gene *CG17604* as the gene *c(3)G*). Simultaneously, Page & Hawley (2001) successively used a construct of $c(3)G^{+}$ and the *gene of green fluorescence protein* to transform mutant *c(3)G* flies and demonstrated localization of C(3)G protein within the synaptic space of pachytene bivalebnt. Thus, our strategy of searching for a *D. melanogaster* TF protein proved to be justifiable.

As soon as for *A. thaliana* and *C. elegans* mutations affecting TF are unknown, the entire genome must be searched. Therefore, in the *A. thaliana* genome, we sought genes that code for proteins similar to Zip1p and SCP1 in domain structure and in length of the alpha-helix in the central domain. Then, the other criteria of protein similarity to Zip1p and SCP1 were employed. We found only one annotated *A. thaliana* gene coding for a protein (AAD 10695) with necessary features (Table). The *C. elegans* genome contains several such genes (according to the information presented in the WormBase and Proteome, Inc. and to our results). On evidence of *in silico* analysis of the structure and putative properties, we chose two proteins, Q11102 and Z81586 (Table), which are potentially able to form TF according to two structural models of SC in *C. elegans*.

Table. Characteristics of experimentally studied and deduced (*) proteins and of SC parameters.

Biological species and SC proteins	Protein (domain) size (amino-acid residues)		SC central space width (nm)	Isoelectric points (pI)			
	whole molecule	Alpha-helix		N-terminal domain	central domain	C-terminal domain	whole molecule
M. musculus SCP1	993	713	100	5,9	5,3	9,7	5,8
S. cerevisiae Zip1p	875	632	115	4,8	6,1	10,1	6,4
D. melanogaster CG17604 *	744	495	109	10,0	4,9	9,7	5,9
A. thaliana AAD10695 *	991	476	100-120	5,3	5,4	9,0	5,6
C. elegans Q11102*	1132	938	70-85	11,9	5,1	11,0	5,5
C. elegans Z81586*	484	460	70-85	4,9	9,5	10,0	9,4

Thus, our strategy allows *in silico* identification of structural proteins that fit the physical parameters and biological properties of subcellular entities with a strongly specified spatial organization. The strategy is best applicable to organisms with known mutations affecting these subcellular structures. When such mutations are unknown, the entire genome must be searched.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (project № 99-04-48182, and 02-04-48761).

References

1. Bogdanov Yu.F., Grishaeva T.M., Dadashev S.Ya. (2002a) Gene *CG17604* of *Drosophila melanogaster* may be a functional homolog of yeast gene *ZIP1* and mammalian gene *SCP1 (SYCP1)* encoding proteins of the synaptonemal complex. Russ. J. Genet. 38, 90-94.
2. Bogdanov Yu.F., Dadashev S.Ya., Grishaeva T.M. (2002b) Comparative genomics and proteomics of *Drosophila*, Brenner's Nematode, and *Arabidopsis*. Identification of functionally similar synaptic genes and proteins. Russ. J. Genet. 38, (№ 8, in press).
3. Heyting C. (1996). Synaptonemal complex: structure and function. Curr. Opin. Cell Biol. 8**,** 389-396.
4. Page S.L., Hawley R.S. (2001) *c(3)G* encodes a *Drosophila* synaptonemal complex protein. Genes Dev. 15**,** 3130-3143.
5. Dong H., Roeder G.S. (2000).Organization of the yeast Zip1 protein within the central region of the synaptonemal complex. J. Cell Biol. 148**,** 417-426.
6. Zickler D., Kleckner N. (1999) Meiotic chromosomes: integrating structure and function. Annu. Rev. Genet. 33, 663-754.

PROTEIN SEQUENCE STUDIES USING FRACTALS

[1] Yenamandra S.P., [2] Mitra C.K.*

Department of Biochemistry, University of Hyderabad, Hyderabad –500 046
e-mail[1]: ckmslrs@uohyd.ernet.in
e-mail[2]: surya_pavan@yahoo.com
*Corresponding author.

Key words: *structure prediction, fractal dimension, fractal interpolation*

Resume

Motivation: Protein structure prediction has become one of the challenging problems in recent times. It has become necessary to know exactly the pattern of distribution of the amino acid residues for predicting the protein structure.

Results: A box counting algorithm has been developed to analyze the dimensional pattern of each individual amino acid in the data bank. Fractal interpolation has been done using fractal dimensions obtained and the preliminary results are graphically presented.

Introduction

Proteins are the most functionally diversified biological molecules. Proteins provide structure, catalyze cellular reactions, act as signaling molecules 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. The amino acids almost never occur in equal amounts in proteins. Although the basic components of all proteins are the same 20 amino acids, proteins function not so much as due to the chemical structure but for the three-dimensional folded structure characteristic of all proteins. Therefore it is important to know the rules that make a particular folded protein structure from a given sequence of amino acids.

Model

We have taken the distribution of amino acid residues in the Swiss Prot protein sequence databank (Release 37, 1999) and calculated the positional distribution of the 20 amino acid residues independently. The graphs appear noisy and no apparent pattern is discernible. Using the box counting algorithm, we have determined the fractal dimensions for all the 20 positional distributions. This was done using a simple program developed by us. These dimensions are later used in the fractal interpolation algorithm. Methionine is anomalous and has not been included in this study.

Brief description of the methodology: We have selected amino acid sequences that are longer than 256 residues long and are not fragments (41,408 sequences). All the computations have been carried out on a PC equipped with a CD-ROM drive, Pentium CPU with 32 RAM. All the necessary soft wares were written in GCC (GNU C++) under Linux operating system. The fractal dimensions of the 20 individual amino acids have been calculated using the Box-Counting Algorithm (Barnsley, 1988a) with minor modifications. Essentially, the dimensions calculated correspond to the Hausdorff-Besicovitch dimension:

$$D = \lim_{n\to\infty} \left\{ \frac{Ln(N_n(A))}{Ln(2^n)} \right\}$$

where $N_n(A)$ is the number of boxes of side length $1/2^n$ that intersect the attractor. For numerical reasons, we cannot take the limit to infinity and have taken the value at n=5 (above this value of n, the set is always discrete). These dimensions are in general agreement with results obtained earlier. Using these values, we have implemented an interpolation algorithm (Barnsley, 1988b). The dimension of the interpolation function is suitably constrained so as to match with the dimension obtained earlier:

$$D = 1 + \frac{Log\left(\sum_{n=1}^{N} |d_n|\right)}{Log(N)}$$

where d_n are the vertical scaling factors (assumed all equal in our case of equally spaced points). To study this algorithm, we have taken 10 representative sequences (manually selected). Protein sequences can be considered as a multi-fractal, with the different amino acid residues showing different fractal dimensions. We notice here that the highest fractal dimension is seen for Tryptophan (W) and the lowest dimension was seen for Alanine (A).

To calculate the fractal interpolation graph, we take a sequence and project all the amino acids independently. The positions of the residue are noted and 20% (64 positions) of the total number of residues (256 was arbitrarily chosen) was considered.

The training set of data points (64 points or every 4th residue) are equally spaced. The interpolation function was selected using an iterative algorithm (Barnsley 1988b) with a constraint that the fractal dimension of the resulting graph is fixed. The interpolated graph is visually compared with the complete distribution.

Results and Discussions

Although fractal dimensions of the distribution of amino acids in protein sequences has been determined earlier (Rani, Mitra, 1996), the interpolation has not been applied for protein sequences so far. The database used 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. The fractal dimensions are calculated up to the fifth level and the fifth level values were considered. The actual values can be read from the Figure below.

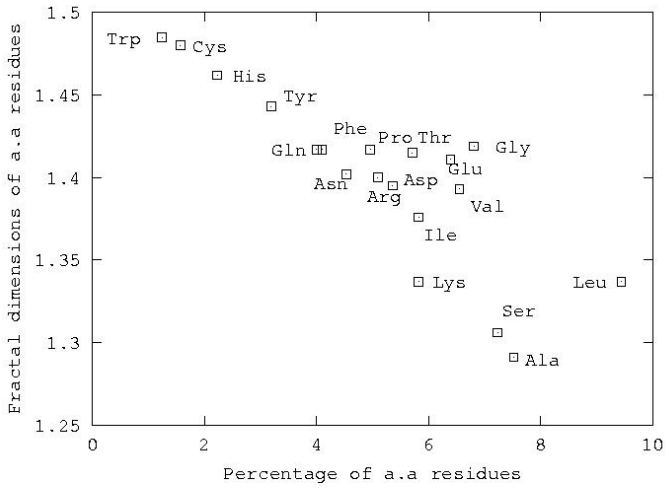


Fig. 1. The fractal dimensions calculated using the box counting algorithm is plotted against its abundance (in the complete database). A negative correlation between these two quantities is clearly seen. However, we can clearly see that the points fall in two distinct categories: Phe, Pro, Thr, Gly, Glu, Val, Leu are somewhat different, at least qualitatively.

For fractal interpolation, we select 64 data points (equally spaced) and use an iterative algorithm. For instance, the sequence we have selected is carboxypeptidase and the residues chosen are Leucine and Phenylalanine. Fractal interpolations of both the residues are clearly seen in the graphs given below (Fig. 2, 3). Leucine has one of the lowest fractal dimension (1.337) and shows less "fractal" character. On the other hand, Phenylalanine (1.417) has higher "fractal" character and this can be seen in the two graphs below.

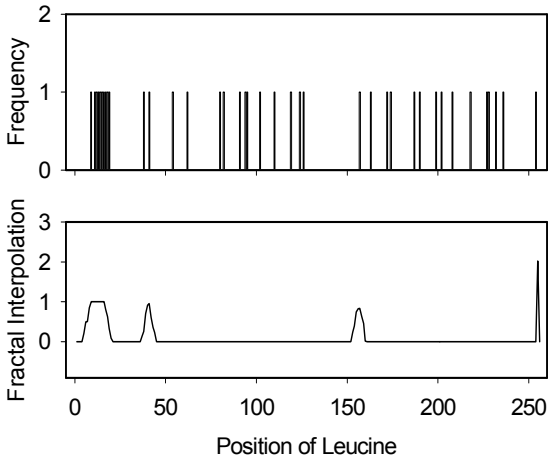


Fig. 2. This shows the distribution of Leucine in the protein sequence "carboxypeptidase" (AC: P10619). As Leucine is generally abundant, we see quite a number of them. The points at y=1 correspond to the presence of Leucine. Similar analysis has been done with all the 20 different amino acids. The lower plot shows the fractal interpolation function for Leucine. This may be directly (and visually) compared with the graph on top that has the actual Leucine residues present. Although not perfect, the similarities are clear. It is to be noted that Leucine has the lowest fractal dimension.

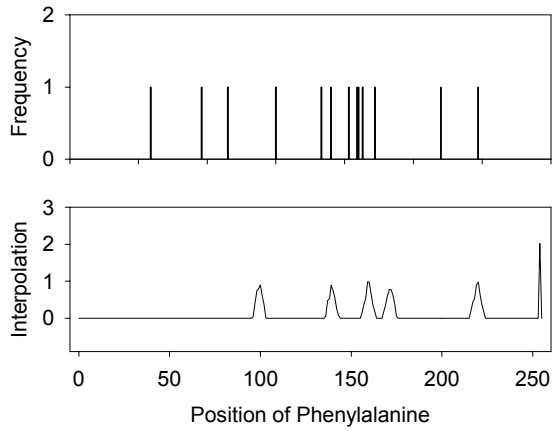


Fig. 3. This shows the distribution of Phenylalanine in the protein sequence "carboxypeptidase" (AC: P10619). Here the fractal interpolation for Phenylalanine can be visualized clearly with the frequency distribution plot of the same.

We have also done the same experiments using the other residues and details of the results will be presented. Program source code is available with authors upon request.

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. (1988a) Fractals Everywhere, Academic Press. 172-206.
2. Barnsley M.F. (1988b) Fractals Everywhere, Academic Press. 207-247.
3. Rani M., Mitra C.K. (1996) Pair-Preferences: a Quantitative measure of Regularities in Protein Sequences. J. Biomolecular Structure and Dynamics. 13, 935-944.

BATMAS30 - THE AMINO ACID SUBSTITUTION MATRIX FOR ALIGNMENT OF BACTERIAL TRANSPORTERS

[1] Sutormin R.A., [2] Rakhmaninova A.B., [1,2] Gelfand M.S.

[1] State Scientific Center GosNIIGenetica, 113545, Moscow, Russia, e-mail: sutor_ra@mail.ru
[2] Integrated Genomics, P.O. Box 348, 117333, Moscow, Russia
*Corresponding author

Key words: *comparative analysis, bacterial transporters, amino acid substitution matrix*

Introduction

Most comparative genomics techniques involve alignment of amino acid sequences and thus depend on amino acid substitution matrices. Therefore it is crucial to develop adequate substitution matrices for different functional regions of proteins.

The best known and the most commonly used substitution matrices are the BLOSUM and PAM series, obtained by statistical analysis of all amino acid sequences. Clearly, in order to align proteins with non-standard physico-chemical characteristics and amino acid composition such as transmembrane proteins, specific matrices are required. The main problem arising during construction of substitution frequency or score matrices for alignment of transmembrane proteins is the fact that in most cases it is not known what part of a protein actually resides within the membrane. The reason is that the transmembrane proteins crystallize poorly, and thus only few such proteins have known spatial structures determined by the X-ray analysis. Different methods for prediction of transmembrane segments yield contradictory results when applied to the same sequence.

We propose the concept of transmembrane kernels (TM-kernels) as a method to find regions of a protein sequence which are most probable transmembrane. We define TM-kernels as parts of the sequences consistently predicted to be transmembrane segments. Two conditions were used: agreement of several prediction algorithms and consistency of predictions for homologous proteins.

Data and Methods

Aligned amino acid sequences of three functionally independent samples of transmembrane (TM) transport proteins have been analized: eubacterial secondary transporters (class TC.2A by the Saier-Paulsen classification (Saier, 1999; Saier, 2000; Paulsen et al., 1998)), eubacterial ABC-transporters (TC.3A.1) and eukaryotic secondary transporters (TC.2A). The basic samples were extended using BLAST homology search over thirty one eubacterial genome in the ERGO system (Overbeek et al., 1993) and eukaryotic proteins from SwissProt.

Every sample was divided into clusters using the nearest neighbor procedure with the percent identity of the BLAST alignment serving as the measure of closeness. When the size of a cluster exceeded 50 proteins, a cluster was further divided into several clusters by raising the lower threshold of clustering. Then each cluster was aligned using CLUSTALW. For each cluster we defined TM-kernels as follows. A position in an amino acid sequence was considered *tentatively transmembrane* (TM-residue) if this position is predicted to belong to a TM-segment by at least three servers out of five: TMHMM (Sonnhammer et al., 1998), TMPRED (Hofmann et al., 1993), DAS (Cserzo et al., 1997), TMAP (Persson et al., 1996), PSORT (Klein et al., 1985). *TM-kernels* in a cluster were defined as groups of adjacent columns in the multiple alignment if each column contains at least 60% of TM-residues. TM-kernels in a protein were defined as groups of positions which belong to the TM-kernel of the cluster. In each cluster, all pairs of sequences with identity in the range ID through ID+10% were considered for varying ID (ID=30%,40%,...,80%). The TM-kernels were used to compute the number of matching amino acid pairs. Then each element of the count matrix was divided by the total of the matrix elements to derive the substitution frequency matrix. Thus, we obtained a series of matrices for different values of ID, named BATMAS (BActerial Transmembrane MAtrix of Substitutions).

Results and Discussion

TM-specific scoring matrices derived using PHDhtm, an algorithm predicting TM-segments in multiple alignment by neural networks, were published by Ng et al. (2000) and Muller et al. (2001). A substitution frequency matrix for highly homologous TM-proteins based on SwissProt annotations was constructed by Jones et al. (1994). Then the Dayhoff mutation model was applied to derive matrices for the comparison of more distant proteins. In all three studies bacterial and eukaryotic proteins were combined into a single sample. The tree in Figure 1 represents the difference between the amino acid composition of several matrices. One can see in this tree that statistical properties of bacterial TM-kernels differs from those of eukaryotic ones. Thus the transmembrane proteins of eubacteria and eukariotes should be considered separately.

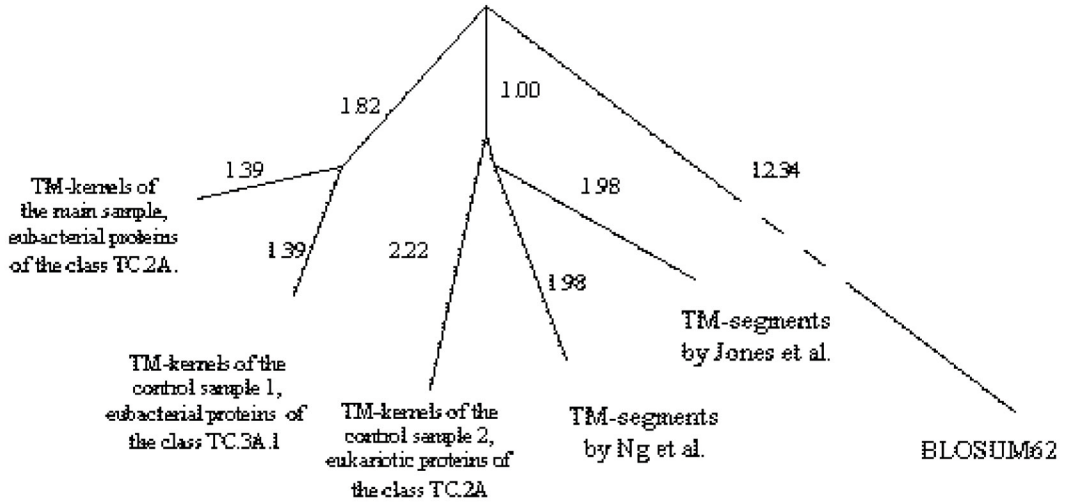


Fig. 1. Comparison of the average amino acid composition of proteins (BLOSUM62) and the amino acid composition of transmembrane segments. The tree constructed by applying UPGMA clustering to the amino acid frequency vectorts.

The average amino acid composition of TM-kernels of the bacterial secondary transporters differs from the published amino acid composition of transmembrane segments in general. TM-kernels contain more alanines, glycines and less polar (cysteine, glutamine), charged (aspartic acid, glutamic acid, lysine, arginine, histidine) and aromatic (tryptophan, tyrosine) residues. The BATMAS30 matrix differs significantly from the standard substitution matrix BLOSUM62, corresponding to the same evolutionary distance: the polar and charged residues as well as proline and tyrosine are highly conserved in BATMAS30 in contrast to BLOSUM62. On the other hand, in BATMAS30 there are more substitutions within the group of hydrophobic residues. Interestingly, tryptophan, which is rare and highly conserved in an average protein, is less conserved in TM-kernels and is more frequently replaced by the polar and charged residues and by proline.

In order to determine the functional role of amino acids, we constructed dendrograms reflecting behaviour of amino acids in TM-kernels. The iterative clustering procedure was as follows: for all pairs of amino acids *i,j* compute $l_{ij}=f_{ij}/(d_i \cdot d_j)$, where f_{ij} is the substitution frequency, d_i is the amino acid probability; merge amino acids *i,j* corresponding to the maximum value l_{ij} into a group and then treat this group as a degenerate amino acid; recompute the substitution frequencies and the amino acid probabilities. Dendrograms constructed for matrices BATMAS30 and BLOSUM62 are shown in Fig. 2. One can see that the topologies of these dendrograms are different. It can be readily seen that in average proteins histidine clusteres with aromatic amino acids, wereas in TM-kernels it is closer to the group of positively charged or polar amino acids. Negatively charged aspartic acid and glutamic acid are form one group in the TM-kernel dendrogram in contrast to the BLOSUM62 matrix where aspartic acid is in the "aspartic" group and glutamic acid is in the "glutamic" group. Cysteine clusters with hydrophobic residues in BLOSUM62, but it is closer to small residues in BATMAS30.

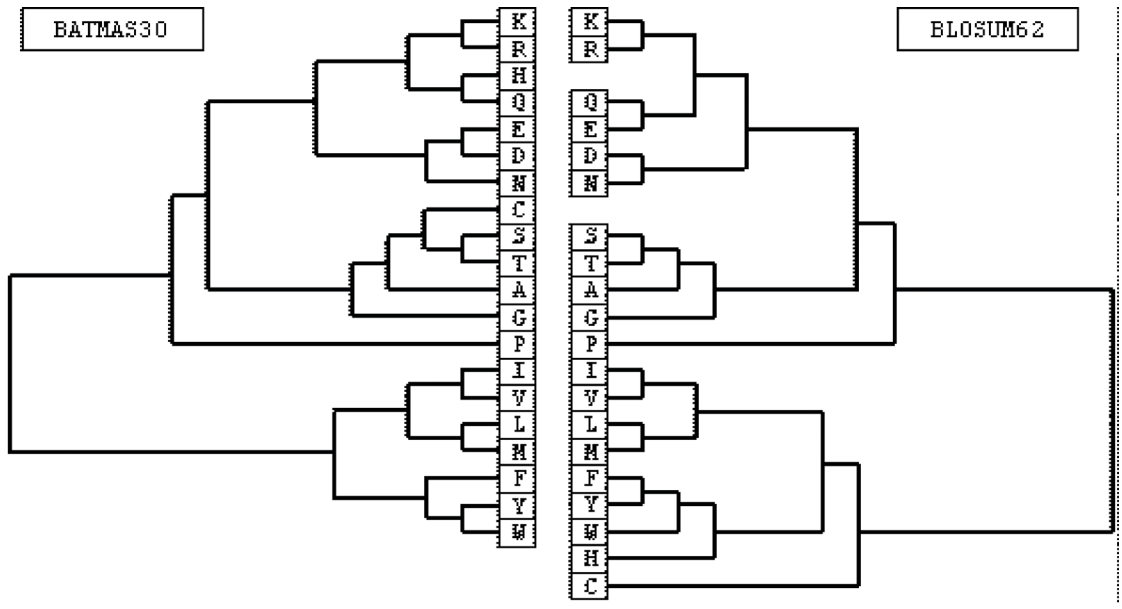


Fig. 2. Dendrograms for the matrices BATMAS30 and BLOSUM62.

Acknowledgments

We thank A.A.Mironov and V.Ju.Makeev for discussion, M.M.Bezruchenkova for assistance with the data set, Maria Chkanikova for translation. This work was partially supported by grants from The Howard Hughes Medical Institute (55000309), INTAS (99-1476) and The Ludwig Cancer Research Institute.

References

1. Cserzo M., Wallin E., Simon I., von Heijne G., Elofsson A. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng. 10:673-676.
2. Hofmann K., Stoffel W. 1993 TMbase - A database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler. 374, 166.
3. Jones D.T., Taylor W.R., Thornton J.M. 1994. A mutation data matrix for transmembrane proteins. FEBS Letters. 339:269-275.
4. Muller T., Rahmann S., Rehmsmeier M. 2001. Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics. 1:182-189.
5. Ng P.C., Henikoff J.G., Henikoff S. 2000. PHAT: A transmembrane-specific substitution matrix. Bioinformatics 16:760-766.
6. Overbeek R., Larsen N., Pusch G.D., D'Souza M., Selkov E.Jr, Kyrpides N., Fonstein M., Maltsev N., Selkov E. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucl. Acids Res. 28:123-125.
7. Paulsen I.T., Sliwinski M.K., Saier M.H.Jr. 1998. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. Mol. Biol. 277:573-592.
8. Persson B., Argos P. 1996. Topology prediction of membrane proteins. Protein Sci. 5:363-371.
9. Saier M.H.Jr. 1999. A functional-phylogenetic system for the classification of transport proteins. Cell Biochem. 32-32:84-94.
10. Saier M.H.Jr. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Mol. Biol. Rev. 64:354-411.
11. Sonnhammer E.L., von Heijne G., Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 6:175-182.

ANCHOR-BASED ALIGNMENT METHOD FOR THE SEQUENCE VS. SEQUENCE AND PROFILE VS. SEQUENCE ALIGNMENT

Sunyaev Sh.R.[1,2], Bogopolsky G.A.[1], Oleynikova N.V.[3,4], Vlasov P.K.[1,3], Finkelstein A.V.[5], Roytberg M.A.[4*]

[1] Institute of Molecular Biology, RAS, Moscow, Russia
[2] European Molecular Biology Laboratory (EMBL), Germany
[3] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russia
[4] Institute of Mathematical Problems in Biology, RAS, Pushchino, Moscow Region, Russia
[5] Institute of Protein Research, RAS, Pushchino, Moscow Region, Russia
e-mail: roytberg@impb.psn.ru
[*]Corresponding author

Key words: *profile, position-dependent scoring matrix (PDSM), alignment*

Resume

Motivation: Sequence vs. sequence and sequence vs. profile alignments are important methods to attribute the protein sequence to the corresponding protein family. However, the accuracy and efficiency of existent methods do not meet the needs of the contemporary genomic and proteomic studies.

Results: The anchor-based method to align protein sequence with another protein sequence or position dependent scoring matrixes (PDSM or profile) was proposed. It has been shown that the method is approximately as accurate as Smith-Waterman method, but considerably faster.

Introduction

Alignment of two protein sequences is old and probably the most classic problem in computational biology. It is a key step in database search, in computational methods for prediction of protein function and homology-based modelling of 3D protein structure. Many sophisticated computational methods in molecular biology, like multiple alignments, profile analysis, threading etc. use pair-wise sequence alignment as a sub-procedure. Smith-Waterman method (Smith, Waterman, 1981) is currently the most sensitive one for alignment, but the slowest. Faster algorithms such as BLAST, FASTA (Altschul et al., 1990, Pearson, 1996) have a tendency to some loss of accuracy.

Using position dependent scoring matrixes (PDSM or profile) usually improves the accuracy and sensibility of alignment. Profile allows to find more distant relevant homologues, because it contains information about multiple or/and structure alignment (Altschul et al., 1990, Eddy, 1998, Sunyaev, 1999).

It is known (Vogt, 1995) that alignments of proteins of low or medium percent of identity (say, 10-30%) obtained by Smith-Waterman method usually differ from those obtained from the 3D-structures alignment. Moreover, our careful comparison between the algorithmic sequence alignments and the structural ones showed that fragments of the Smith-Waterman alignments with pour similarity usually have nothing to do with the structural (and thus more reliable) alignment.

This suggests to ignore (at least for beginning) all elements of the Neadleman-Wunsch matrix, except the “anchors”, i.e. ungapped fragments of (relatively) high similarity.

The idea to start the alignment procedure from the search for such anchors is definitely not new and was implemented in several software tools (e.g. mentioned above BLAST and FASTA). However, this idea was considered as a way to increase computational speed of alignment techniques inevitably associated with the loss of alignment accuracy. Our observations suggest the way to improve computational speed without sacrificing (and even with a slight gain of) alignment accuracy and confidence compared to the Smith-Waterman algorithm. The technique is adopted both for sequence vs. sequence and for sequence vs. profile alignments.

Methods

The proposed alignment algorithm works up given sequences in three steps:

Step 1. We generate a set of ungapped high-scoring segments ("*anchors*"), marked as shadow cells on Fig. a.

Anchor is an ungapped matching of equal-length fragments, {U[a, a+L] vs. V[b, b+L]}, of sequences U and V. These fragments meet the following conditions:

a) anchor contains at least one “seed pair” {U[x, x+1] vs. V[y, y+1]} with the score exceeding a cutoff CSeed;

b) the anchor’s score (i.e., the sum of the substitution scores M(U[x], V[y]) over the anchor) exceeds a cutoff CAnchor;

c) the score of any continuous part of the anchor exceeds a cutoff CMin;

d) the anchor is "locally maximal", i.e., (i) it is not a part of any other pair of segments {U[a', a'+L'] vs. V[b', b'+L']} meeting conditions a) – c) and having greater or equal score, and (ii) it does not include any continuous part having a greater score.

Step 1. starts with identifying seed pairs and then expands them to obtain the anchors. This step is similar to the procedures used in BLAST and FASTA and is the most time-consuming step of our algorithm.

Step 2. We find the optimal *block alignment* path through the set of anchors (marked by thick line). Block is a continuous part of anchor. Block alignment is a chain of the blocks $\{B_1, \ldots, B_N\}$, where B_i precedes B_{i+1} both along U and V sequences. The block alignment $\{B_1, \ldots, B_N\}$ is optimal if it has maximal possible *block score,* which is defined as follows:

$$\text{Score}(B_1, \ldots, B_N) = \text{Score}(B_1) - \text{Link}(B_1,B_2) + \text{Score}(B_2) - \ldots - \text{Link}(B_{N-1},B_N) + \text{Score}(B_N). \quad (1)$$

$\text{Score}(B_i)$ is the total score of matches along block B_i according to the given substitution matrix M. $\text{Link}(B_i, B_{i+1}) = \alpha + \beta \bullet |(y-x) - (y'-x')|$ is the linkage penalty for the blocks B_i and B_{i+1}, where α (*linkage open penalty)* and β (*linkage elongation penalty)* are analogs of the traditional gap opening and gap elongation penalties, while x, y are the last residues of block B_i, and x', y' are the first residues of block B_{i+1} in sequences U and V, respectively. Note that we penalize links between the blocks even if the blocks are placed on the same diagonal.

To find the optimal block alignment from the created set of anchors we use either the Wilbur-Lipman algorithm (Wilbur, Lipman, 1983) (if the number of the initial anchors K is small), or (if K is large) the sparse dynamic programming (SDP) method (Eppstein et al., 1992). These procedures produce the same alignments (given the same parameters and set of anchors), but differ in the run-time: the Wilbur-Lippman algorithm run-time is proportional to K^2, while the SPD run-time is of order K*log(L), where L is the length of the shorter sequence. The first procedure performs faster (and therefore is used to find the optimal block alignment) if $K < 20$; otherwise the second procedure is evoked.

Step 3. We specify the alignment path in regions between the blocks. To this end we use a global version of the Smith-Waterman algorithm. Our experiments show that usually this step comprises only a small part of the total run-time of our algorithm.

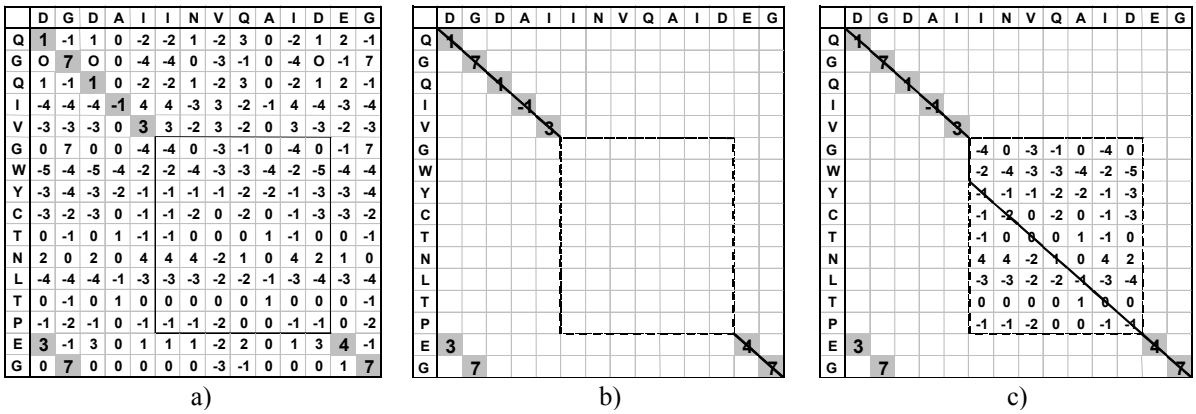


Fig. Three steps of new alignment algorithm: (a) finding of anchors set; (b) obtaining of the optimal block alignment path through the set of anchors; (c) improving the alignment between anchors by a global version of the Smith-Waterman algorithm.

The main difference between implementations of the method for the sequence vs. sequence and sequence vs. profile alignment is in creation of anchors (step 1). In both cases we start with the generation of all possible seed pairs with the similarity score exceeding a cutoff Cseed. To do this we scan the first sequence or the profile U. For each position *i* of U we create the list of all k-tuples (usually, k = 2) having significant (> CSeed) similarity with the k consequent positions of U, starting in *i*-th position. The run-time for this step is proportional to the length of U, but if U is a profile, the multiplicative constant is larger. This is the most time consuming step of the algorithm. Fortunately, in case of the database search this step is performed only once per the whole of the database and therefore its run-time is not crucial for the effectiveness of the database search. When the lists of the potential homologues of the k-tuples are created, the standard FASTA-like technique to generate the anchors can be implemented.

Results and Discussion

Accuracy and confidence of the method have been tested through comparison with 583 standard alignments extracted from BaliBase databases (Thompson, Plewniak, Poch. Bioinformatics, 1999, 15, 87-88). For each pair of protein sequences the golden standard alignments have been compared to the alignments constructed by Smith-Waterman algorithm with standard settings. Percentage of amino acid pairs correctly aligned by an algorithm with respect to the golden standard alignment was

used as a measure of the algorithmic alignment ***accuracy*** and percentage of amino acid pairs correctly aligned by an algorithm with respect to the algorithm alignment was used as a measure of the algorithmic alignment ***confidence***.

For testing profile method we use the following technique: from the database of multiple alignment we delete one sequence, produce profile [PSIC] and align the deleted sequence with obtained profile.

Results of these tests show that the novel method slightly outperforms the SW algorithm both in accuracy and confidence. As has been stated above, focusing on high-scoring regions can significantly improve the computational speed of the algorithm. The current software has been compared with the standard (search at http://biobase.dk/programs/Ordered_by_Funtionality/Multifunctional_Programme_Pack/PEARSON/Pearson_Programme_List/pearson_programme_list.html) and in house implementations of the SW algorithm. Table shows that the suggested method requires about 1.5 times shorter computational time than the classic SW technique. Origins of these issues discussed in the Methods section.

Table. Presents the data on the characteristics of the algorithm for align protein sequences.

Family	Num	Len	%ID	KOP_time	SW_Time	SW_Acc%	SW_Conf%	KOP_Acc%	KOP_Conf%
1idy	10	54	12.7	9	16	15.66	28.96	13.01	28.88
1tvxA	21	58	21.1	12	19	24.92	40.61	25.12	41.26
1aboA	120	59	24.9	8	20	65.02	72.07	64.44	71.21
1tgxA	105	63	37.0	16	23	61.86	68.37	63.21	67.90
1r69	10	70	15.5	15	28	23.84	33.98	23.84	30.35
1ubi	6	85	24.2	22	42	28.15	51.36	32.56	52.56
1csy	55	85	30.8	19	41	67.23	69.93	68.79	70.35
2trx	6	92	18.8	27	47	27.30	39.23	27.30	39.23
1wit	10	97	17.4	25	55	43.85	66.44	43.14	56.66
1uky	6	203	15.0	152	235	29.63	35.97	33.55	35.44
1havA	171	211	29.9	130	249	58.41	69.99	60.31	66.41
2hsdA	6	245	18.8	209	344	39.20	43.00	36.44	39.42
1sbp	15	245	16.1	210	341	13.71	14.47	16.62	17.28
2pia	6	257	13.2	202	379	47.47	60.49	47.87	58.51
kinase	6	270	24.0	276	412	71.00	75.35	73.48	78.01
1ped	3	350	24.7	684	704	53.89	65.25	52.68	62.61
4enl	3	364	20.3	738	768	28.74	30.03	31.3	32.26
1ajsA	6	371	13.7	724	813	15.41	27.14	25.32	28.78
1cpt	6	398	20.8	853	946	61.90	70.47	62.39	67.15
2myr	6	407	15.8	1082	997	20.60	21.66	22.20	23.00
1pamA	6	470	20.5	1499	1377	45.03	57.89	50.55	53.08
All	***583***	***138***	***27.6***	***111***	***162***	***55.07***	***63.79***	***56.20***	***62.30***

Data for logarithmic data set of parameters for pare-wise alignments: 15 (opening) and 1 (extension) deletion penalties for Smith-Waterman algorithm; CAnchor = linkage open penalty = 17 and linkage elongation penalty = 1; substitution scoring matrix = Gon250. For profile vs. sequence alignments data is the same.

Notation: family – protein's family name in Bali-base; Num – number of performed alignments; Len – average length of sequences in the family; %ID – average percent of identity; KOP_time and SW_time – time need to make new and SW alignment respectively; _Acc% and _Conf% - average percent of accuracy and confidence.

Acknowledgements

This work was supported by the INTAS grant 99-01476, by the Netherlands Organization for Scientific Research (NWO) grant, by 00-04-48246, 01-04-48400, and 01-01-00287 RBRF grants, 13/hg grant from Russian State programme Human Genome, partly by French-Russian Lyapunov Centre and by an International Research Scholar's Award to A.V.F. from the Howard Hughes Medical Institute.

References

1. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25:3389-3402
2. Eddy S.R. (1998) Profile hidden Markov models. Bioinformatics. 14, 755-763.

3. Eppstein D., Galil Z., Giancarlo R., Italiano G.F. (1992) Sparse dynamic programming. 1. Linear cost-functions. J. of the ACM. 39, 519-545.
4. Pearson W.R. (1996) Effective protein sequence comparison. In Meth. Enz., R.F.Doolittle, ed. (San Diego: Academic Press). 266:227-258
5. Smith T.F., Waterman M.S. (1981) Identification of common molecular subsequences. J. Mol. Biol. 147, 195-197.
6. Sunyaev S.R., Eisenhaber F., Rodchenkov I.V., Eisenhaber B., Tumanyan V.G., Kuznetsov E.N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 12(5):387-394.
7. Vogt G., Etzold T., Argos P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zone revisited. J. Mol. Biol. 249, 816-831.
8. Wilbur W.J., Lipman D.J. (1983) Proc. Natl Acad. Sci. USA. Rapid similarity searches of nucleic acid and protein data banks. 80, 726-730.

NEW METHOD OF LATENT PERIODICITY DETECTION MAY DETERMINE STRUCTURALLY RELATED PROTEINS AND PROTEIN FAMILIES

[1] Laskin A., [1,2*] Korotkov E., [2] Kudryashov N.

[1] Bioengineering Center, RAS, 117312, Moscow, Russia
[2] Moscow Physical Engineering Institute, 115409, Moscow, Russia
e-mail: katrin2@biengi.ac.ru
[*]Corresponding author

Key words: ***periodicity, alignment, regular structures, nucleotide-binding domains, repeat***

Resume

Motivation: The study of repeats in biosequences is important because repeats in sequence often lead to repeats or regularities in structure. Recently we elaborated novel methods allowing locating subsequences in genetic texts that consist of hidden (non-homologous) repeats. But since the only criterion for the presence of latent periodicity (the property of a sequence composed of hidden repeats) is its statistical significance, we need to know whether other related sequences have this type of periodicity and thereby this periodicity pattern has some structural or functional sense.

Results: We proposed and implemented the algorithm for cyclic alignment, which is a locally optimal alignment of a sequence against a cyclically elongated profile. Insertions and deletions are allowed in both sequence and profile, giving the algorithm more sensitivity to ancient and distant repeats where they are likely to arise. In addition, the profile can be trained to search for structurally and functionally related periodicity types. In our analysis of Swiss-Prot we found a number of previously unknown periodicity patterns corresponding to structural features of NAD-binding sites of dehydrogenases, active sites of protein kinases, dethiobiotin synthetases and others. For instance, a 24 residues long periodicity pattern turned out to correspond to Rossman-fold (or Rossman-like) domains. The analysis of periodicity profile showed that it contains alteration of α and β-structure preference, and this alteration is known to be common for nucleotide-binding domains.

Availability: The software (C++ source and executables for different platforms) and detailed results for investigated types of periodicity are available upon request from the authors.

Introduction

The search for distant evolutionary correlation between protein and DNA sequences has promoted the notions of enlarged similarity (Korotkov, Korotkova, 1996) and latent (hidden) periodicity (Korotkov, Korotkova, 1995; Chaley et al., 1999) in amino acid and nucleotide sequences. Latent periodicity can be imagined as enlarged similarity to some perfectly periodic sequence. Since we can find enlarged similarity in many cases where homology searches fail, latent periodicity also represents more than just the occurrence of tandem homologous repeats. The definitions of both notions are based on an information-statistical metric instead of conventional Sellers metric. In 1999 we found latent periodicity in about 10% of sequences in the Swiss-Prot data bank (Korotkov et al., 1999; Korotkova et al., 1999). It was shown that in many cases we can observe periodic regularity in an unequal distribution of symbols in a sequence where traditional homology-based search methods failed to detect any repeats. We think that the main difference between these approaches and our one is that they are concentrated on findings of homologous subsequences (repeats) while we try to investigate the whole sequence.

For many proteins we found latent periodicity in sites that were not even supposed to be composed of repeats or duplications, so we made a proposition that such regularities may reflect subtle signals of their evolution, conserved mostly on the structural level. We suppose that many of functional protein domains arose by duplications of relatively short sequences, which had ability to interact with substrates or coenzymes, as is the case for zinc-finger domains, and then evolved in conjunction to form even more functional domain together.

Methods and Algorithms

We intended to search for tandem repeats with insertions and deletions, assessing statistical significance by simulation and extrapolation, where necessary. Our technique is based on techniques developed in (Fischetti et al., 1992), but adopted to search for optimal subsequence (or subsequences) with latent periodicity since we concentrated our investigations on domains that we considered as having internal hidden periodic structure. For mathematical accuracy we invented the cylindrical coordinate space with one (cyclic) coordinate corresponding to position in cyclic profile and another (linear) coordinate corresponding to conventional sequence position. We introduced cyclic alignment as a path on that cylindric surface and its score in usual way, summing the weights of symbols at corresponding positions in cyclic profile and

subtracting the costs of all insertions and deletions that we made in that alignment. For locally optimal alignment we did not take into account the starting and ending deletions as usual.

We found that we can solve the problem of finding of the optimal cyclic alignment and its score using well-known dynamic programming technique, adopted for working in cylindrical coordinate space. Namely, we introduce the partial similarity matrix $S_{i,j}$, where index *i* corresponds to wrapped coordinate, i.e. position in the periodic pattern (all calculations of it are to be performed modulo L, where L is the period length) and index *j* is conventional position in a sequence. This matrix can be filled as follows:

If optimal alignment exists (not necessarily uniquely), then the following interrelation holds:

$$S_{i,j} = \max\left\{S_{i-1,j-1} + w_{i,j}, \max_{1\le k\le j}\left\{S_{i,j-k} - d_k\right\}, \max_{1\le k\le L-1}\left\{S_{i-k,j} - d_k\right\}\right\}$$

where $w_{i,j}$ is weight of symbol at position j in the pattern position i and d_n is the penalty of insertion/deletion of n successive symbols. If we are to construct a locally optimal alignment, this formula has to be slightly rewritten:

$$S_{i,j} = \max\left\{0, S_{i-1,j-1} + w_{i,j}, \max_{1\le k\le j}\left\{S_{i,j-k} - d_k\right\}, \max_{1\le k\le L-1}\left\{S_{i-k,j} - d_k\right\}\right\}$$

The main difference is that i is now a cyclic coordinate. This leads to recursive interdependence between values of $S_{i,j}$, so they cannot be calculated using the formulae above. But this interdependence can be resolved this way:

Denote $S'_{i,j} = \max\left\{0, S_{i-1,j-1} + w_{i,j}, \max_{1\le k\le j}\left\{S_{i,j-k} - d_k\right\}\right\}$.

Then, assuming that $d_m + d_k \ge d_{m+k}$, $S_{i,j} = \max\left\{S'_{i,j}, \max_{1\le k\le L-1}\left[S'_{(i-k),j} - d_k\right]\right\}$.

Furthermore, if the gap penalties are given in affine form, i.e. $d_n = a + b*n$, there is an obvious way to reduce the number of calculations and thereby the running time of our algorithm (Gotoh, 1982).

To assess the statistical significance of alignments we used the Monte-Carlo method. We performed a given number of alignments with shuffled sequences for each highly positive score separately. The distribution of alignment scores was found to be close to normal, so we computed mean E(S) and variance D(s) of these random scores and obtained Z-value using the formula:

$$Z = \frac{S_{real} - E(S)}{\sqrt{D(S)}}$$

Sufficiently high Z-values (Z>6.0) were treated as evidence for latent periodicity in the investigated sequence.

Periodicity profiles for initial searching were obtained from the results of our previous studies, where a method of searching for latent periodicity without deletions was proposed. In addition, we used structure information to determine the true period length for structurally related cases. After the initial search we were able to rebuild the profile using the information obtained from aligned sequences. We obtained the aggregate alignment matrix of all significant periodicity cases. From this matrix we recalculated the new profile according to the formula:

$$w_{i,j} = \ln \frac{N_{i,j}}{f_j \sum_{all\ k} N_{i,k}}$$

This could be done as many times as we wanted, but usually this process had to be stopped after a number of iterations due to stabilizing of number of found periodic sequences or increasing of false positives rate.

Results and Discussion

The algorithms for searching for significant cyclic alignment and profile optimization were implemented in the software suite that is capable of identifying structure-related sites in protein sequences. The software was written in C++ for x86-compatible computers running Windows, as well as for clusters of x86-compatible computers running Linux. The running time of the Swiss-Prot scanner depends greatly on the period length and the number of performed significance tests; it ranges from several minutes to several hours on a single PC.

Swiss-Prot data bank release 39 with updates was used for searching. We also used special selections from Swiss-Prot during profile training to avoid possible false positives at this stage (for example, to optimize the profile for NAD-binding sites search we used a subset of Swiss-Prot entries marked with "NAD" keyword). Many latent periodicity profiles obtained in our previous investigations (Korotkov et al., 1999; Korotkova et al., 1999) were tested; below is a brief review of the most impressive results we achieved.

To illustrate the sensitivity of our methods to identify known homologous repeats we used 28 residues long profile obtained from C2H2-type zinc finger protein. After optimization we identified all 315 entries with 3 or more C2H2-type zinc-finger domains (we used 3 times the period length as a cut-off for identifying a latently periodic subsequence) at very high significance level (Z~16-70). No false positives were found. In many cases the periodic subsequence exceeded the range of

tagged zinc-finger domains so we can say that we can identify more zinc-fingers in those sequences than conventional Pfam or PROSITE-based search.

We investigated the 24 residues long profile obtained from structural alignment of some NAD-binding sites. When optimized it was capable of finding total 2196 proteins with corresponding type of periodic structure and 1330 of them (~60%) were tagged as interacting with NAD or its analogs. The major part of the remainder was denoted as interacting with other nucleotides (FAD, ATP, GTP etc.). These nucleotide-binding domains are known to be structurally similar and characterized by alteration of α and β-structures. We investigated the profile and found that it consists of α-structure and β-structure preference sites.

Previously we also obtained periodicity with period length of 18 residues in a few protein kinases. We trained this profile using two subsets of Swiss-Prot: one with serine-threonine protein kinases and another with tyrosine protein kinases, so we obtained two periodicity profiles. The resulting serine-threonine profile was capable of detecting of 774 out of 920 serine-threonine protein kinases and tyrosine profile was capable of detecting of 301 out of 326 tyrosine protein kinases. Most of total 60 false positives were caused by another type protein kinases; only 8 of them were not protein kinases at all.

We also investigated 30 residues long profile initially obtained from dethiobiotin synthetase. Using it we found all 8 known dethiobiotin synthetases with no false positives. Dethiobiotin synthetases have torus-like 3D structure and a period corresponds to a turnover; it also consists of α-structure and β-structure preference sites.

Conclusions

Our studies confirm that latent periodicity is the structure- or function-related feature of primary protein sequences. Our method allows effective searching for latently periodic sites in sequence data banks and correlating them to known protein families.

References

1. Chaley M.B., Korotkov E.V., Skryabin K.G. (1999) Method revealing latent periodicity of the nucleotide sequences for a case of small samples. DNA Res. 6, 153-163.
2. Fischetti V., Landau G., Schmidt J., Sellers P. (1992) Identifying periodic occurrences of a template with applications to a protein structure. In Apostolico A. et al. (eds), Proc. of the 3rd annual Symposium on Combinatorial Pattern Matching, Lecture Notes in Computer Science volume 644, Springer-Verlag, 111-120.
3. Gotoh O. (1982) An improved algorithm for matching biological sequences. J. Mol. Biol. 162, 705-708.
4. Korotkov E.V., Korotkova M.A. (1995) DNA regions with latent periodicity in some human clones. DNA Seq. 5, 353-358.
5. Korotkov E.V., Korotkova M.A. (1996) Enlarged similarity of nucleic acid sequences. DNA Res. 3, 157-164.
6. Korotkov E.V., Korotkova M.A., Rudenko V.M., Skryabin K.G. (1999) Latent periodicity regions in amino acid sequences. Mol. Biol. 33, 611-617.
7. Korotkova M.A., Korotkov E.V., Rudenko V.M. (1999) Latent periodicity of protein sequences. J. Mol. Model. 5, 103-115.

RARE RESIDUES FORM THE CHANNEL IN TRANSMEMBRANE TRANSPORTERS

[1] Kalinina O.V., [1] Makeev V.Ju., [1] Sutormin R.A., [1,2] Gelfand M.S., [2] Rakhmaninova A.B.

[1] State Scientific Center GosNIIGenetica, Moscow, 113545, Russia
[2] Integrated Genomics, P.O. Box 348, Moscow, 117333, Russia
e-mail: linel@mail.ru
*Corresponding author

Key words: *membrane proteins, bacteria, transporters, statistical analysis, channel prediction*

Resume

Motivation. Transmembrane transport is of extremely importance for the cell life. Many genes encoding real or putative transport proteins are found in bacterial genomes. But in most cases their substrate specificity is not experimentally determined and only approximately predicted from genomic studies. Even less is known about the

3D-structure of transporters. Nevertheless the published experimental data lets us assume that determination of the channel-forming residues would allow make suppositions about substrate specificity of secondary transporters.

Results. We have developed a simple computational method for identification of channel-forming residues in transporter sequence, based on our original data about amino acids frequencies in bacterial secondary transporters. We have applied it to transmembrane proteins with resolved 3D structure and the prediction showed a sufficiently good agreement with the real protein structure.

Availability. All results are available from the authors (linel@mail.ru).

Introduction

Transmembrane (TM) transporter proteins are the major mechanism of the flow of compounds in and out the bacterial cell. Up to eleven percent of a prokaryotic proteome are the membrane transporter systems, and thus prediction of their substrate specificity not only is important for the genome annotation, but also is of major practical interest. The experimental data, though scarce, indicates that in the case of secondary transporters, the substrate specificity is determined by the general structure of the TM channel. Therefore, prediction of the substrate specificity requires determination of the channel-forming residues.

Although only few resolved 3D structures of transporters are known, there are many structural models based as on various indirect experimental data. However, different prediction algorithms yield contradictory results when applied to the same sequence, and the same algorithm may yield contradictory results when applied to orthologous proteins. In an accompanying abstract (Sutormin et al.) we introduce the concept of TM-kernels as a protein fragment consistently predicted to be a transmembrane segment.

The aim of this study was to develop a method for identification of channel-forming residues using statistical analysis of TM-kernels.

Statistical analysis of TM-kernels

We have analyzed 18908 kernels from 2172 proteins (bacterial secondary transporters, class 2.A according to the Saier-Paulsen classification). The TM-kernels retain the periodic distribution of residues described for the whole TM-helices.

To reveal amino acid residues propensity to lie on the same or on the opposite sides of a TM-helix we calculate ***positional correlation for groups of amino acid residues***.

Let M be the number of TM-kernels in the sample. Let l_k be the number of residues (length) of k -th kernel. Consider two disjoint groups of residues, α and β. The positional correlation for each distance *n* was calculated as follows. Let $N_n^{\alpha\beta}$ be the number of such residue pairs, where the first residue belongs to group α, the second residue belongs to group β) and the distance between the residues is (*n-1*): $N_n^{\alpha\beta} = \sum_{k=1}^{M}\sum_{i=1}^{l_k} I^{\alpha}(x_i) I^{\beta}(x_{i+n})$, where $I^{\alpha}(x) = \begin{cases} 1, x \in \alpha \\ 0, x \notin \alpha \end{cases}$.

Let N_n be the number of all pairs at the distance (*n-1*). $N_n = \sum_{k=1}^{M} (l_k - n)$.

Finally, let p_α be the frequency of residues from group α in the sample of TM-kernels: $p_\alpha = \sum_{k=1}^{M}\sum_{i=1}^{l_k} I^\alpha(x_i) \Big/ \sum_{k=1}^{M} l_k$.

Then the positional correlation coefficient in point n is $\mathrm{corr}(n) = \dfrac{N_n^{\alpha\beta} - N_n p_\alpha p_\beta}{\sqrt{p_\alpha(1-p_\alpha)p_\beta(1-p_\beta)} \cdot \sum_{k=1}^{M} l_k}$.

We have observed that tryptopran and tyrosine tend to lie at the same side of the helix as charged and polar residues (Fig. 1). We assume that this is the channel side and therefore call them *channel residues*. The common property of these residues is that according to our data (Sutormin et al.) their frequency in TM-kernels is significantly less than in protein in general. Still, the average number of channel residues per kernel is 2.6 (Fig. 2), which might be sufficient for determination of the channel side of a helix.

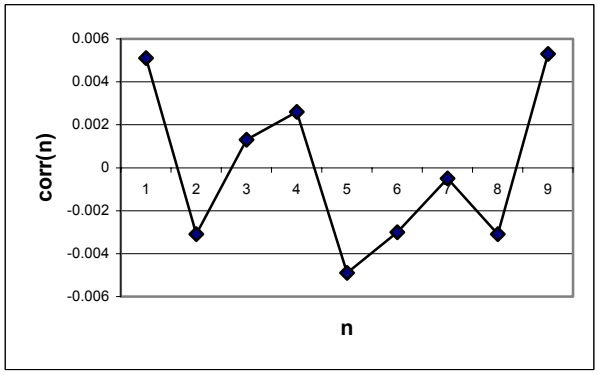


Fig. 1. Positional correlation between two groups of amino acids: charged (K, R, H, Q, D, E, N) and aromatic (F, W, Y) amino acids.

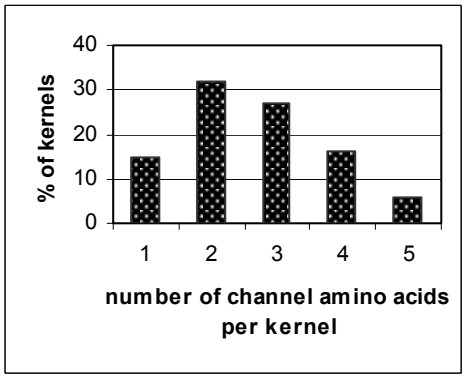


Fig. 2. Distribution of the number of channel amino acid residues in kernels.

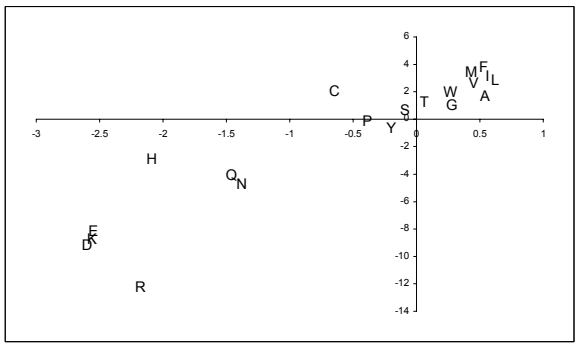
Calculation of the channel moment

Two scales of channel propensity were constructed as follows:

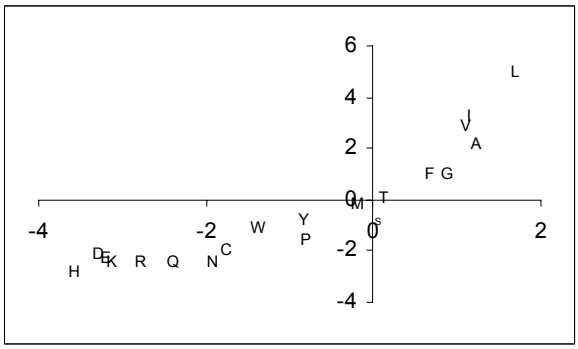
$$P^{(1)}{}_a = \log\frac{f_a^{\mathrm{tm}}}{f_a^{\mathrm{av}}}, \qquad P^{(2)}{}_a = \log\frac{f_a^{\mathrm{tm}}}{1/20},$$

where $P^{(\nu)}{}_a$ is the channel propensity of residue *a*, f_a^{tm} is the frequency of *a* in TM-kernels, f_a^{av} is the frequency of *a* in all proteins.

Correlation of these scales with about 90 different scales of amino acid attributes [http://pref.etfos.hr/split/] used for prediction of TM helices was computed. As expected, $P^{(1)}$ turned out to be similar (correlation coefficient >0.85) to several scales but there still are some numerical differences. The other scale, $P^{(2)}$, correlates with only one scale (Fig. 3a, b). It must be noted that both scales showed only poor correlation with most popular scales, such as the Kyte-Doolittle scale (Kyte, Doolittle, 1982) ($P^{(1)}$ and $P^{(2)}$: 0.84), Eisenberg scale (Eisenberg et al., 1984) ($P^{(2)}$: 0.79) and kPROT (Pilpel et al., 1999) ($P^{(1)}$: 0.46, $P^{(2)}$: 0.48).



a. Correlation of $P^{(1)}$ (horizontal) with Engelman scale (Engelman et al., 1986) (vertical) (correlation coefficient = 0.93).



b. Correlation of $P^{(2)}$ (horizontal) with the Kuhn-Leigh scale (Kuhl, Leigh, 1985) (vertical) (correlation coefficient = 0.90).

Fig. 3. The published scales having the highest correlation with the channel propensity scales [http://pref.etfos.hr/split/].

These scales were used for identification of *channel residues*. ***The channel moment C*** was defined analogously to the hydrophobic moment (Eisenberg et al., 1984):

$$\vec{C} = \sum_i \vec{c_i} ,$$

where $\vec{c_i} = \vec{r_i} \cdot P^{(\nu)}{}_a$, $\vec{r_i}$ is the radius-vector of residue at position *i*, $P^{(\nu)}{}_a$ is the channel propensity scale #ν (ν = 1, 2).

Several bacterial and archaeal TM proteins with resolved 3D structure were used to test the reliability of rotational orientation of TM-segments by the channel moment.

To determine the "true" orientation of the channel vector we calculated the vector pointing to the most exposed side of the helix and assumed that it points to the membrane, that is, ***out*** of the channel. That was done using the solvent accessibility surfaces from the DSSP database [http://www.sander.ebi.ac.uk/dssp/] or calculated using the program SPDBV [http://cn.expasy.org/spdbv/]. We considered only proteins which had an inner cavity or channel and an easily detectable single layer of helices surrounding this cavity: 1FBB (bacteriorhodopsin, *Halobacterium salinarum*), 1E12 (light-driven chloride pump, *Halobacterium salinarum*), 1H68 (sensory rhodopsin II, *Natronomonas pharaonis*), 1FX8 (glycerol-conducting channel, *Escherichia coli*), 1MSL (mechanosensitive ion channel MSCL homolog, chain A, *Mycobacterium tuberculosis*), 1BL8 (KCSA, potassium channel, chain A, *Streptomyces lividans*). In the latter case the outer helices were removed from the PDB file. Visual control and analysis of positions of functionally important residues showed that this procedure adequately describes the channel. Total number of TM helices in the study was 32.

Results

The angle differences between the calculated channel moments and the directions of "true" channel vectors for all 32 studied TM helices are shown in Fig. 4. One can see that the obtained predictions are comparable to the ones obtained using the Kyte-Doolittle scale: in approximately 2/3 cases the channel side is predicted with deviation less than 50° from the "true" direction and in about 1/3 cases the channel side is predicted badly. The latter phenomenon is coupled with the objective limit of accuracy for such predictions: some helices contain charged residues that face the membrane, possibly establishing interactions between protein subunits.

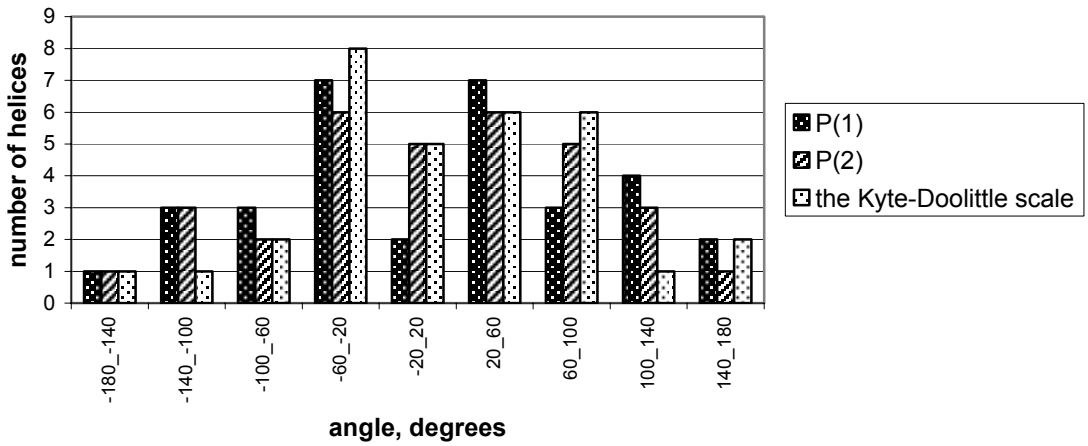


Fig. 4. Comparison of different scales for orientation of TN-helices relative to the channel. Horizontal axis: the angle between the channel moment and the true channel direction.

Additionally, we have analyzed MsbA (Chang, Roth, 2001), which is the only bacterial transporter with resolved 3D structure. The numerical analysis is impossible since the X-ray structure of MSBA is still incomplete: only coordinates of Cα-atoms are published. The visual analysis reveals good accuracy of our predictions: We have analyzed the residues that lie within the sector of 90° facing the predicted channel direction (±45° from the channel moment) and all but three of them indeed face the channel. Moreover, all six residues, shown in (Chang, Roth, 2001) to face the channel, lie in the predicted sector.

The most intriguing result of this study is that predictions done using $P^{(2)}$ are not worse than those obtained by any other scale. It means that one can predict the channel side of a TM-helix without prior assumptions about the amino acids properties and using only amino acid frequencies in TM-kernels.

We plan to use this method for modeling secondary transporters. We expect a good accuracy of prediction, especially for monomeric proteins.

Acknowledgements

This study was partially supported by grants from RFBR, INTAS (99-1476), HHMI, and LICR/CRDF. We are grateful to A.A.Mironov for useful discussion.

References

1. Chang G., Roth C.B. (2001) Structure of MsbA from *E. coli*: a homolog of the multidrug resistance ATP binding cassette (ABC) transporters. Science. 293: 1793-1800.
2. Eisenberg D., Schwarz E., Komarory M., Wall R. (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. 179: 125-142.
3. Engelman D.M., Steitz T.A., Goldman A. (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Ann. Rev. Biophys. Chem. 15: 321-353.
4. Kuhl L.A., Leigh J.S. (1985) A statistical technique for predicting membrane protein structure. Biochim. Biophys. Acta. 828: 351-361.
5. Kyte J., Doolittle R.F. (1982) A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157: 105-132.
6. Pilpel Y., Ben-Tal N., Lancet D. (1999) kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. J. Mol. Biol. 294:921-935.
7. Sutormin R.A., Rakhmaninova A.B., Gelfand M.S. BATMAS30 - the amino acid substitution matrix for alignment of bacterial transporters. This volume.

RECOGNITION OF OCCURRENCE AND LOCALIZATION OF CLEAVAGE SITE IN SIGNAL PEPTIDES

*[1] Zagoruiko N.G., [1] Kutnenko O.A., [2] Nikolaev S.V., **[2,3] Ivanisenko V.A.

[1] Institute of Mathematics, SB RAS, Novosibirsk, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[3] State Research Center of Virology and Biotechnology "VECTOR", Koltsovo, Novosibirsk region, Russia
*e-mail: zag@math.nsc.ru, ** salix@bionet.nsc.ru

Key words: *signal peptide, signal anchor, cleavage site, informative characteristics, recognition*

Resume

Motivation: Automatic recognition of signal peptides and cleavage sites in proteins is a topical issue for both detection of their cellular localization and solving of applied medical and biotechnological problems. The available recognition methods utilize either amino acid substitution matrices (von Heijne, 1986) or neuronal network algorithms using a 20-letter amino acid code (Nielsen et al., 1997).

In this work, the feasibility of using physicochemical characteristics of amino acids for recognizing cleavage sites in signal peptides is studied. The algorithm AddDel (Zagoruiko, 1999) was applied for selecting the significant characteristics. Cleavage sites were recognized by a sliding test technique. The rule of "k nearest neighbors" at $k = 1$ was used as a decisive rule.

Results: A method for selecting the informative subset of characteristics was developed and the decisive rule based on the membership function was constructed. Testing of the method proposed using a large amount of experimental data has demonstrated that it detects the cleavage sites correctly and localizes at a rate of 85%.

Introduction

Signal peptides are N-terminal markers of proteins. Eukaryotic proteins carrying these markers are transported through the endoplasmic reticulum membrane; prokaryotic, through the internal membrane. Upon transportation, signal peptidase cleaves the signal peptide. The signal peptides display a common structure: a short positively charged N-terminal region, central hydrophobic region, and a more polar C-terminal region, containing the site whereat the peptide bond is cleaved. X-ray structure analysis of signal peptidase demonstrated that the spatial matching of the enzyme in question and a signal peptide required that the amino acid residues at position −1 and −3 with respect to the cleavage point were small (von Heijne, 1998). N-Terminal regions of type II membrane proteins are signal anchors, providing the integration of these proteins with the membrane. In their physicochemical properties, signal anchors are close to signal peptides; however, unlike the latter, they are not cleaved by signal peptidases (Sakaguchi et al., 1992).

The goal of this work was to study the feasibility of using physicochemical characteristics of amino acids for recognizing signal peptides and signal anchors. The signal peptides were recognized according to the presence of cleavage sites using the method of k nearest neighbors.

Methods and Algorithms

Forming the learning sample. The learning sample used was represented by three sets of fragments of eukaryotic proteins, namely, (1) the fragments containing cleavage sites, (2) the fragments containing anchors, and (3) the fragments of nuclear and cytoplasmic containing neither sites nor anchors. All the necessary data were extracted from http://www.cbs.dtu.dk/services/SignalP/.

Physicochemical properties of amino acids. Two set of amino acid characteristics were used: (1) a Kidera's set of 10 properties (Kidera et al., 1985), that is, noncorrelating linear combinations of amino acid structural and physicochemical properties, and (2) 434 structural and physicochemical properties from the database http://www.genome.ad.jp/dbget/aaindex.html.

Selecting the window size and decisive rules. At the first stage, we studied the accuracy of cleavage site recognition depending on the window size, that is, on the number of symbols taken into account to the left and right of the cleavage point. The Kidera's set of properties was used as a character; the window size was changed from 6 to 36 symbols.

At this stage, the sample was formed of fragments with a length L containing cleavage site at the center. The negative sample was formed of a random protein fragments lacking the cleavage site. Thus, a data array was formed of representatives of two images as a table comprising 1012 lines (objects) and L × 10 columns (characters).

A method of sliding test was used was used for recognition; method of k nearest neighbors, as the decisive rule. The limit window size of 18 symbols was selected due to a limited data volume; all the further studies were performed with this window size.

Selecting the descriptors. The informative amino acid characteristics were selected from each set of physicochemical properties. As the size of the negative sample exceeded considerably that of the positive sample, 7 learning samples containing the same 253 elements from the positive sample and 7 different sets of 253 elements each from the negative sample were formed. The data for the control recognition were not used in the learning samples.

Seven most informative characteristics were selected from the Kidera's set of ten properties (properties Nos. 3, 5, and 10 were excluded) using the window of a size L = 18. The reliability of recognition in the learning sample using these characteristics amounted to 87.6%.

Then, the informative characteristics were selected out of the 434 physicochemical properties plus the 10 Kidera's properties. The algorithm AddDel (Zagoruiko, 1999) was used for this purpose. This algorithm combines the concepts of "successive addition of most valuable"(Addition) and "successive deletion of least valuable"(Deletion) properties. It appeared that comparatively small number of properties—from 7 to 30—gave the best results. Overall, 91 property of 444 tested were included into these 7 sets. The Kidera's properties failed to display their advantages—only one character was included into one of the sets.

These seven sets formed of 444 characteristics were selected as sets of descriptors.

Recognition in the control sequence. The window with a size of 18 symbols was moved along the protein chain with a shift of 1 symbol. Overall, 296,202 control regions were distinguished by this technique; of them, 252 fragments contained the cleavage site and 295,950 were without the site. The decision on the presence or absence of the cleavage site was made at each window position. These two images were recognized according to the seven sets of characteristics described above in parallel. The decision in favor of either first or second image was made by a majority of these seven votes.

When 252 fragments containing the cleavage site were subjected to recognition, the correct decisions rate amounted to 213 or 84.5%. The number of correct decisions while studying the fragments lacking the site amounted 232,403 (78.5%).

Localization of the cleavage site. Estimation of the probability of the presence or absence of the cleavage site in the sliding window was calculated using a modified rule of the k closest neighbors. The distances r1 and r2 to two closest neighbors, one from each image, were found for each control object y. The function of membership in a certain image was specified as $f = 1 - 2*r1/(r1 + r2)$. The value of function f changes in the range of –1 to +1. If $f \geq 0$, then the object y belongs to the first image; in the opposite case, to the second image. The overall value of the function F was obtained by a mere averaging of the membership functions f obtained for each descriptor set. The mean reliability of recognizing the two images appeared equal to 82.5%.

Recognition of the signal peptides containing cleavage site and of signal anchors. The technique developed was applied to solving the problem of recognition of the three images: Signal Peptide, Anchor, and Negative (fragments of cytoplasmic and nuclear peptides). The signal peptides were recognized according to the presence of cleavage site; the signal anchor, according to the C-terminal boundary of the membrane region in protein. The method of pairwise comparison (Zagoruiko, 2002) was used for recognizing these images. For each pair of images, an individual "competent" space, where these images differ maximally from one another, was formed. Information on the frequencies of all the symbols of the alphabet used at the first two odd positions of the window was used in addition to physicochemical properties. While performing the pairwise comparisons, the solution was made according to the value of the membership function. The overall decision on the membership of control object in one of the three images was made basing on the results of the pairwise comparison.

Results and Discussion

Symmetric and asymmetric windows of various sizes were analyzed. It appeared that the window with four odd symbols, two at either side from the center, gave the best results. Thus, the window of size *L* equaling eight symbols was used in the further studies.

Localization of cleavage site. The cleavage site recognition function averaged over 252 control proteins containing the cleavage site is plotted in Fig. 1. The protein fragments are positioned with respect to the cleavage site position (vertical line). The plot in Fig. 2 demonstrates the same function obtained for 252 randomly selected cytoplasmic and nuclear proteins lacking the cleavage site. The length of the fragments amounted to 84 symbols. It is evident from Figs. 1 and 2 that the cleavage site is recognized and localized well.

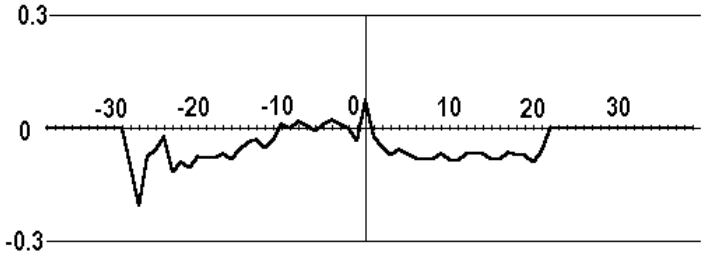


Fig. 1. Cleavage site recognition functions for the fragments containing the site.

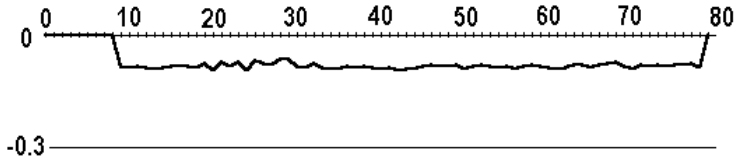


Fig. 2. Cleavage site recognition functions for the fragments lacking the site.

Recognition of the signal peptides containing cleavage site and of signal anchors.

The results of recognizing signal peptides and signal anchors are listed in Table 1. The control sample comprised 65,607 objects, including 705 signal peptides, 47 signal anchors, and 64,855 objects of the negative sample. These results demonstrate a satisfactory capability of discriminating between signal peptides and signal anchors.

Table. Recognition of signal peptides and signal anchors.

Presented	Recognized		
	Signal peptide	**Anchor**	**Negative**
Signal peptide	**549**	50	106
Anchor	14	**25**	8
Negative	11873	12540	**40442**

The experiments on recognizing signal peptides and signal anchors have demonstrated that the additional information on amino acid frequencies at certain position within the fragments increases the accuracy of recognition.

As an example, plots of changes in type I and II errors versus the threshold value of the membership function *F* while recognizing signal anchors (47 objects) with and without the data on amino acid frequencies are shown in Fig. 3. The negative set comprised 50,111 objects.

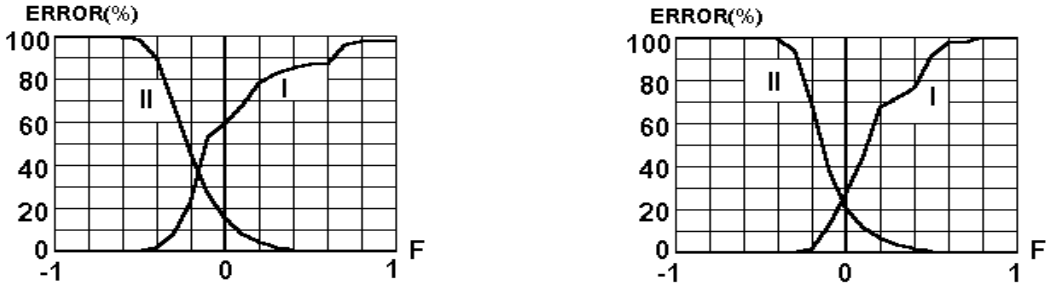


Fig. 3. Type I and II errors while recognizing signal anchors versus the threshold value of membership function considering (**a**) only physicochemical amino acid properties and (b) supplementary data on their frequencies.

Testing the "oddness" hypothesis. Significance of the positions even and odd with respect to the cleavage site while its recognition was studied. The symbols were numbered from left to right starting from 0 to L-1. The experiments have demonstrated that positions with odd numbers contribute to the recognition to a greater degree, thereby complying with the known (–1; –3) rule (von Heijne, 1998).

Acknowledgements

The work was supported by Russian Foundation for Basic Research (grants № 02 01-00082 to N.G.Zagoruiko and O.A.Kuntenko and № 01-07-90376 to V.A.Ivanisenko and S.V.Nikolaev) and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65 Simulation of Basic Genetic Processes and Systems).

References

1. von Heijne G. (1986). A new method for predicting signal sequence cleavage sites. Nucl. Acids Res. 14, 4683-4690.
2. von Heijne G. (1998). Life and death of a signal peptide. Nature. 396, 112-113.
3. Kidera A., Konishi Y., Oka M., Ooi T., Scheraga H.A. (1985) Statistical analysis of the physical properties of the 20 naturally occurring amino acids. J. Prot. Chem. 4, 23-55.
4. Nielsen H., Engelbrecht J., Brunak S., von Heijne G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering. 10, 1-6.
5. Sakaguchi M., Tomiyoshi R., Kuroiwa T., Mihara K., Omura T. (1992). Functions of signal and signal-anchor sequences are determined by the balance between the hydrophobic segment and the N-terminal charge. Proc. Natl Acad. Sci. USA. 89, 16-19.
6. Zagoruiko N.P. (1999). Applied Methods for Data and Knowledge Analysis. Novosibirsk: Izd. Instituta Matematiki.
7. Zagoruiko N.P. (2002). Image recognition by the method of pairwise reference comparison in competent subspaces of characters. Dokl. Akad. Nauk. 382, 24-26.

COMPARISON OF METHODS FOR PREDICTING PROTEASOME CLEAVAGE MOTIFS

[1] Nikolaev S.V., [1] Afonnikov D.A., [1,2] Ivanisenko V.A., [2] Bazhan S.I., [1] Kolchanov N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: nikolaev@bionet.nsc.ru
[2] State Research Center of Virology and Biotechnology "VECTOR", Koltsovo, Novosibirsk region, Russia

Key words: *proteasome, proteolysis, proteasome cleavage motifs, antigenic epitope, MHC1, algorithm, prediction*

Resume

Motivation: Proteasome proteolysis is one of the major processes of degradation of proteins with structural defects or antigens. Methods for predicting proteasome cleavage motifs are of great importance for medicine and biotechnology. At present, there are several algorithms of prediction of such motifs. In the present paper, we have chosen three of them for comparison.

The algorithm by Kesmir et al. (2002) is based on experimental data on proteasome cleavage *in vitro*. The other two algorithms (Altuvia, Margalit, 2000; Kesmir et al., 2002) are based on data on peptides binding to the major histocompatibility complex (MHC) class 1. In this study, we compare the predictions of proteasome cleavage motifs performed with these three algorithms.

Results: We have shown that the results of the test can be interpreted as follows: two algorithms (Kesmir et al., 2002) detect complementary features of proteasome cleavage motifs. Combination of their predictions has improved the recognition of the motifs in the test sample of fragments.

Introduction

Degradation of proteins in the cell is vital for both the regulation (e.g., rapid digestion of transcription factors) and elimination of defective proteins and antigens (Ciechanover, 1998). Proteasomal cleavage is one of the main processes of such degradation. It is performed by a protein complex, proteasome. The core of the proteasome is a multienzymatic complex, whose subunits are endoproteases integrated into a cylindrical aggregate consisting of four rings. Each of the rings consists of seven subunits. This allows cleavage of an enzyme next to any amino acid residue but in a certain context (Orlowski et al., 2000). It is believed that peptides resulting from this cleavage have the same C terminus as the MHC1 antigenic epitope (Altuvia, Margalit, 2000). The N terminus of the peptide can be additionally modified by aminopeptidases in the cytosol and/or cytoplasmic reticular lumen (Altuvia, Margalit, 2000).

Recent experimental (Nussbaum et al., 1998) and theoretical (Holzhutter et al., 1999; Altuvia, Margalit, 2000) studies of proteasome cleavage have brought about a number of methods for prediction of cleavage motifs. The present paper is dedicated to testing the three algorithms providing quantitative estimates of cleavage site scores.

Methods and Algorithms

Prediction algorithms

Algorithm 1 (Kesmir et al., 2002) (NC20S). A neural network trained with experimental data on *in vitro* proteolysis. Available via *http://www.cbs.dtu.dk/services/NetChop/*, network type '20 S'.

Algorithm 2 (Kesmir et al., 2002) (NCC2). A neural network trained with data on MHC1-related peptides. Available via *http://www.cbs.dtu.dk/services/NetChop/,* network type 'C-term 2.0'.

Algorithm 3 (AM). The Cleavage Scores Table calculated with the use of data on MHC1-binding peptides (Altuvia, Margalit, 2000), is available by the address *http://bioinfo.md.huji.ac.il/marg/cleavage/scores.html*. We used this table for predicting cleavage motifs as follows. A sequence was scanned with a window with a length equaling two amino acid residues (A, B). Cleavage site scores were selected from the Cleavage Scores Table at the cross of row A and column B and were assigned to the first residue of the window. The scores picked from the Cleavage Scores Table were normalized to the range [0, 1]. The algorithm is implemented in the form of a script in the system Matlab 5.2.0.

Comparison of Algorithms

The algorithms for predicting cleavage motifs were compared using a sample of protein fragments with the proteasome cleavage motifs determined *in vitro* (Holzhutter et al., 1999). The sequences of these fragments are shown in Fig. 1.

1) insulin B chain
Fv**NQHLC**g**SHLVEALYLVCGERGFFYTPK**a
2) viral peptide HBVcAg
Ayrpp**NAPILST**lpe**TTVVRR**r**GRSP**rrr**TP**s
3) viral peptide pp89
r**LMYDMY**phf**M**ptn**LG**ps**EK**r**V**w**M**s
4) OvaY 51-71
Yqt**IN**k**VVRFD**k**LP**g**FGD**si**E**a
5) OvaY 249-269
YVsg**LE**q**LE**si**INFE**k**L**t**E**wts
6) Ova 239-281
ms**ML**v**LL**pde**VS**gl**EqLESI**n**FE**k**L**t**EWtSS**n**VM**ee**RKI**kvyl
7) p53wt
Tle**D**ssgn**LL**g**R**ns**F**e**V**r**VC**acpgrdr

Fig. 1. Sequences used for testing the algorithms for predicting cleavage motifs. Capitalized are amino acid residues next to which the proteasome proteolysis was experimentally observed *in vitro* (Holzhutter et al., 1999).

For all the tested sequences, proteolysis motif scores were either obtained from www servers or calculated (for algorithm 3). Then, all the sequences and their experimental and predicted cleavage motifs were pooled into one sample.

Probabilities of type I and II errors were calculated in this sample for various threshold levels. The dependence curves for type I and II (false positive and false negative) errors are shown in Fig. 2.

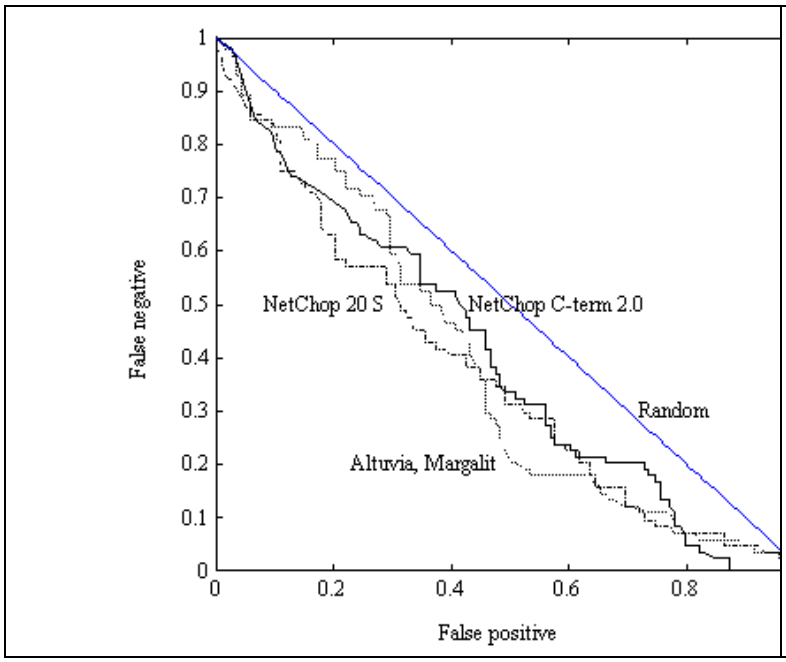
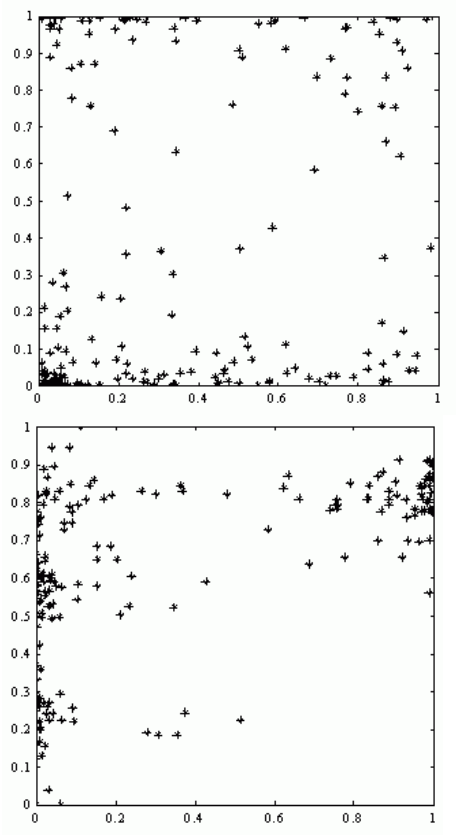


Fig. 2. The plot of the dependences of false negative errors on false positive errors for three prediction methods: solid line, NetChop C-term 2.0; dotted line, Altuvia and Margalit; and dot–dash line, NetChop 20 S.

We verified the agreement between the predictions of cleavage motifs made by these algorithms for various positions of the test sequences. For this purpose, we constructed scattering diagrams of the predicted scores for each algorithm pair (Fig. 3).

Probabilities of type I and II errors were calculated for combined predictions. Predictions made with the threshold of 0.5 were taken into account. The following combinations were considered: (1) disjunction of predictions, (2) conjunctions of predictions, (3) disjunction of conjunctions, and (4) "voting" method (Table).



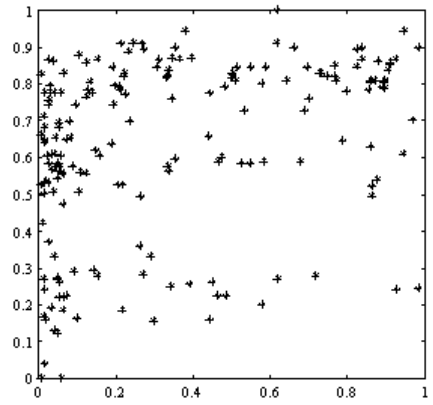


Fig. 3. Plots of cleavage motif scores: (a) X-axis: NetChop 20 S, Y-axis: NetChop C-term 2; (b) X-axis: NetChop 20 S, Y-axis: Altuvia & Margalit; (c) X-axis: - NetChop C-term 2, Y-axis: Altuvia and Margalit.

Table. Probabilities of errors of types I and II for combined predictions.

		Type I error	Type II error	Sum of type I and II errors
1.	NetChop C-term 2.0 (NCC2)	0.59	0.24	0.83
2.	NetChop S 20 (NC20S)	0.62	0.21	0.83
3.	Altuvia and Margalit (AM)	0.22	0.72	0.94
4.	$NCC2 \cup NC20S$	**0.41**	**0.36**	**0.77**
5.	$NCC2 \cap NC20S$	0.81	0.1	0.82
6.	$NCC2 \cap AM$	*0.6*	*0.24*	*0.84*
7.	$NC20S \cap AM$	*0.65*	*0.24*	*0.89*
8.	$(NCC2 \cap NC20S) \cup (NCC2 \cap AM)$	0.6	0.24	0.84
9.	$(NCC2 \cap NC20S) \cup (NC20S \cap AM)$	0.65	0.17	0.82
10.	$(NCC2 \cap AM) \cup (NC20S \cap AM)$	***0.45***	***0.31***	***0.76***
11.	***Voting method***	***0.45***	***0.31***	***0.76***

Results and Discussion

The plots of cleavage motif prediction errors (Fig. 2) show that the algorithms tested have approximately equal ratios between the errors of types I and II. Note that in the range of type I errors from 0.2 to 0.4, the least errors of type II occur in the case of the algorithm NC20S, trained on experimental data.

Figures 3a–c show that the cleavage motif scores predicted by different algorithms are in poor agreement. This may result from the fact that one algorithm was trained only on experimental data on *in vitro* proteasome proteolysis, and the two other algorithms, on the C-ends of antigenic epitopes. The last two algorithms show a certain agreement in predictions: dots tend to occur in the left upper corner of Fig. 3c. In addition, the algorithm AM yields overprediction in comparison with NCC2, probably, because the former uses positions P1 and P2 for evaluation of the scores, and the latter uses a wider window around the peptide bond examined.

Probabilities of prediction combination errors A_i made by the tested algorithms are shown in Table. The following combinations were considered: (1) disjunction of prediction pairs $(A_i \cup A_j)$, (2) conjunction of prediction pairs $(A_i \cap A_j)$, (3) disjunction of conjunction pairs $(A_i \cap A_j) \cup (A_k \cap A_l)$, and (4) voting method $(A_1 \cap A_2) \cup (A_1 \cap A_3) \cup (A_2 \cap A_3)$. The

probabilities of correct prediction for test fragments increase when predictions by the methods NCC2 and NC20S are joined (row 4 vs. rows 1 and 2, Table). This may indicate that the training samples for the neural networks NCC2 and NC20S have complementary features, and their weight matrices also code for complementary information. This is confirmed by the low probability of the conjunction of predictions by these networks (row 5, Table). In addition, Fig. 2 shows that mere threshold change cannot yield this ratio between errors of types I and II by none of these methods.

Conjunction of predictions made by AM with those made by NC20S and NCC2 (rows 6 and 7, respectively) somewhat reduces the prediction level but less than with algorithms 1 and 2. Disjunction of these conjunctions also increases the prediction probability (row 10, Table) and is as good and the combined prediction by the "voting" method (row 11, Table).

For all methods, the threshold level of 0.5 is taken. Disjunction of predictions by the methods NetChop C-term 2 и NetChop 20 S increases the probability of correct prediction (row 4).

Thus, our comparison has demonstrated that the available methods for predicting proteasome cleavage motifs are sensitive to the training sample. Disjunction of the predictions made by the neural networks NetChop C-term 2 and NetChop 20 S improves the motif recognition in the test sample of fragments.

Further prediction improvement will depend on both developing the models underlying the prediction methods and obtaining new experimental data. Note that methods for comparison of algorithms for predicting proteasome cleavage motifs with a limited volume of experimental information should also be developed.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); US Department of Energy (grant № 535228 CFDA 81.049); and CRDF (grant № RB0-1276).

References

1. Altuvia Y., Margalit H. (2000). Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. J. Mol. Biol. 295: 879-890.
2. Ciechanover A. (1998). The ubiquitin-proteasome pathway: on protein death and cell life. EMBO J. 17: 7151-7160.
3. Holzhutter H.G., Frommel C., Kloetzel P.M. (1999). A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20 S proteasome. J. Mol. Biol. 286: 1251-1265.
4. Kesmir C., Nussbaum A., Schild H., Detours V., Brunak S. (2002). Prediction of proteasome cleavage motifs by neural networks. Prot. Eng. 15: 287-296.
5. Nussbaum A., Dick T.P., Keilholz W., Schirle M., Stevanovic S., Dietz K., Heinemeyer W., Groll M., Wolf D.H., Huber R., Rammensee H.G., Schild H. (1998). Cleavage motifs of the yeast 20S proteasome b subunits deduced from digests of enolase 1. Proc. Natl Acad. Sci. USA. 95: 12504–12509.
6. Orlowski M., Wilk S. (2000). Catalytic Activities of the 20S proteasome, a multicatalytic proteinase complex. Arch. Biochem. Biophys. 383: 1-16.

AN INDEX FOR ESTIMATING THE EFFICIENCY OF ANTIGENIC EPITOPE GENERATION DURING PROTEASOMAL PROTEOLYSIS

[1] Nikolaev S.V., [1,2] Ivanisenko V.A., [1] Afonnikov D.A., [2] Bazhan S.I., [1] Kolchanov N.A.*

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: nikolaev@bionet.nsc.ru
[2] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk region, Russia
* Corresponding author

Key words: *proteasome, proteolysis, antigenic epitope, MHC-1, prediction*

Summary

Motivation: Proteasomal proteolysis is one of the major processes involved in presentation of antigenic epitopes related to the major histocompatibility complex class 1 (MHC-1). Evaluation of the efficiency of epitope processing during proteasomal proteolysis is of paramount importance for medicine and biotechnology; in particular, for prediction of antigenic determinants and development of polyepitope vaccines.

The goal of this study was to develop an index of the efficiency of epitope processing during proteasomal proteolysis basing on the data on location of cleavage sites.

Results: An index of the efficiency of epitope processing and a method for its calculation from the location of proteasome cleavage sites are proposed. The index has been applied to estimation of the yield of a known epitope with alanine substitution mutations in the epitope flanking regions.

Introduction

Antigenic epitopes of MHC-1 launch cell-mediated immune response. They emerge in the following chain of intracellular processes. The antigen molecule is ubiquitinated and degraded by a proteasome, which is a multienzyme complex. A proteasome core is a cylinder consisting of four rings, each of which is made up of seven subunits. This provides the proteolysis next to any amino acid unit in a protein but in a certain context (Ciechanover, 1998; Orlowski et al., 2000). Epitopes are certain parts of peptides emerging during proteolysis. It is believed that the C-terminus of a peptide enters an epitope, whereas the N-terminus can be truncated by aminopeptidases in the cytosol and/or cytoplasmic reticular lumen (Altuvia, Margalit, 2000).

Peptides displaying affinity for the transporter molecule associated with antigen processing (TAP) are transported into cytoplasmic reticular lumen. Peptides with affinity for MHC-1 are associated with it and form the epitope–MHC-1 complex. This complex is presented on the cell surface (Lauvau et al., 1999).

Recently, a number of algorithms for prediction of cleavage sites for proteasomal proteolysis have been developed (Holzhutter et al., 1999; Altuvia, Margalit, 2000; Kesmir et al., 2002). It has also been shown that the accuracy of cleavage site prediction can be increased by combination of such predictions (Nikolaev et al., 2000).

We propose a method for calculating the efficiency of epitope processing during proteasomal proteolysis based on cleavage motif scores.

Methods and Algorithms

Epitope generation index

The following expression is proposed as an index of the yield of an epitope. It reflects the proportion of antigen molecules that yield fragments containing the epitope after proteasomal proteolysis:

$$Ef(epitop) = \frac{\sum_{\forall i,j,(i,j)\supseteq epitop} S(i,j)}{\sum_{\forall i,j,(i,j)\cap epitop\neq 0} S(i,j)}, \qquad (1)$$

where $S(i,j)$ is the score of the fragment (i,j). The numerator is the sum of scores of all the fragments (i,j) containing a particular epitope, and the denominator is the sum of scores of all fragments (i,j) intersecting with this epitope. If the denominator is null (no proteolysis motifs yield fragments intersecting with this epitope), then $Ef(epitop) = 0$.

Calculation of the score of a fragment *S*(*i*,*j*)

To calculate the score of a fragment, we have considered cleavage of a protein sequence according to two models.

The first model assumes that proteolysis at individual positions within a sequence occurs independently (Kesmir et al., 2002). Then, the probability of the occurrence of a fragment starting at position *i* and ending at position *j* can be calculated as follows:

$$p(i,j) = p(i) \times \overline{p}(i+1) \times ... \times \overline{p}(j-1) \times p(j), \quad (2)$$

where $p(l)$ is the probability of endoproteolysis, and $\overline{p}(l)$ is the probability of the absence of endoproteolysis at position *l*. By taking the logarithm of Equation (2), we obtain its additive form. By analogy with it, we have an index of a fragment score, where scores of motifs have no probabilistic interpretation but are additive:

$$S(i,j) = S(i) + (S_{\max} - S(i+1)) + ... + (S_{\max} - S(j-1)) + S(j), \quad (3)$$

where $S(i,j)$ is the score of a fragment represented in terms of cleavage site scores; $S(k)$ is the score of a cleavage site according to a certain scoring system.

According to the other model, the efficiency of proteolysis at position (*i*) depends on whether proteolysis has occurred at position (*j*) (Holzhutter et al., 1999). In this case, the expression

$$S(i,j) = S(i) + S(j) + \alpha \times \ln(p(L_{ij})) \quad (4)$$

can be chosen as fragment score. Here, $S(k)$ is the score of the motif in a certain scoring system, $p(L_{ij})$ is the *a priori* (experimental or predicted) frequency of fragments of the length L_{ij}, and α is a normalization factor for the chosen system of motif scores. In this model, if $S(i)$ or $S(j)$ equal zero, then $S(i, j)$ is also zero.

Implementation and Results

To illustrate the application of the proposed index, we have compared the efficiency of the generation of the antigenic epitope K L L E P V L L L of the 40S ribosomal protein (Rammensee et al., 1999). Replacement of substitutions in positions flanking the epitope has been modeled: Q48 → A, Y49 → A and G59 → A, K60 → A. Profiles of cleavage motif scores predicted by the program NetChop (*http://www.cbs.dtu.dk/services/NetChop/*) are shown in Fig. 1. The network type is 'C-term 2.0'.

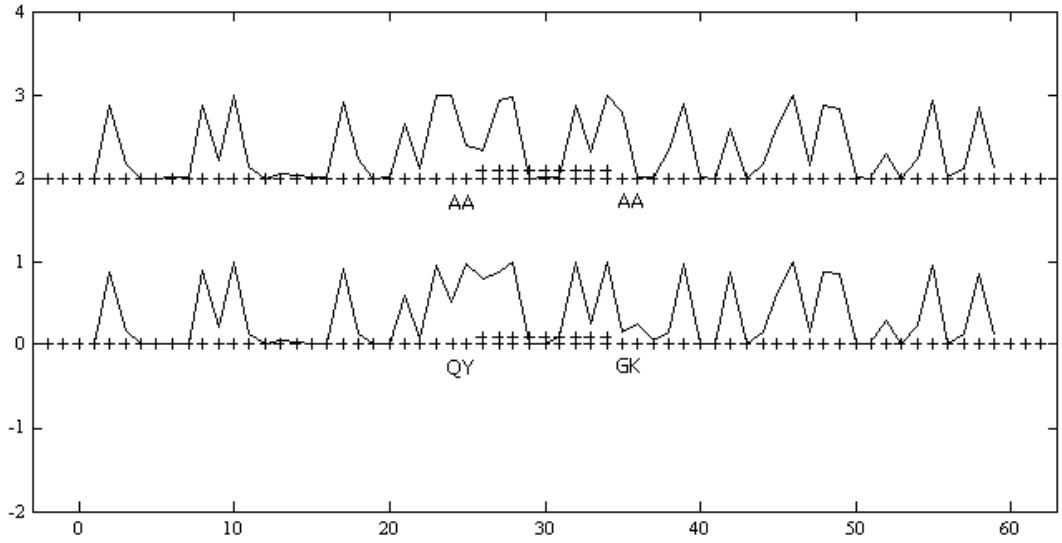


Fig. 1. Profiles of cleavage motif scores predicted by the program NetChop. The epitope domain is marked with double plus signs ‡. Lower plot: profile for the native protein fragment 25–83. Upper plot: profile for the mutant protein.

Figure 1 shows that the mutation decreases the cleavage motif score in the mutant protein immediately before the N terminus of the epitope and increases after the C terminus. This would reduce the probability of the epitope generation. On the other hand, the cleavage motif score is also reduced at position –3 from the C end of the mutant fragment. This increases the score of the fragment (with proper ends) (Fig. 2).

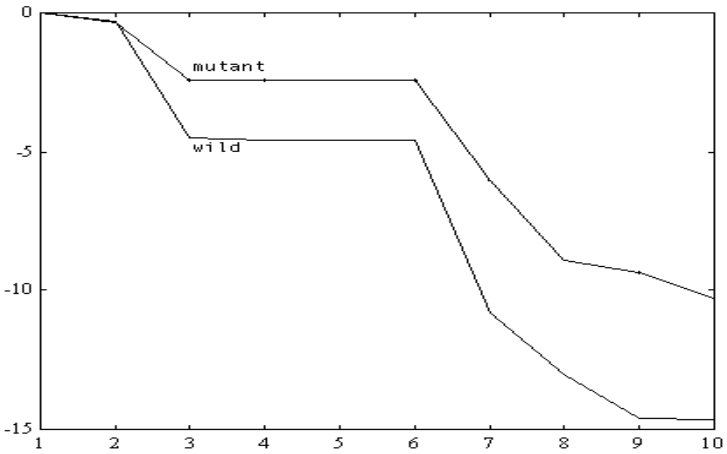


Fig. 2. Score of a fragment versus its length. The length is measured from the C end of the fragment (according to the concept of the role of proteasomal proteolysis in the formation of the N and C ends of antigenic epitopes).

Such variations in fragment scores have opposite effects on the efficiency of the epitope generation. Thus, it is difficult to guess the resulting variation in the epitope yield.

The efficiency of fragment generation calculated by Equation (3) is 0.0072 for the native protein and 0.0040 for the mutant. Logarithms of the results obtained by the program NetChop are taken as motif scores.

Thus, the model of independent proteolysis motifs assumes (*ceteris paribus*) that the antigenic activity of this epitope within the native protein is higher than in the mutant one.

Discussion

It is known that mutations in the vicinity of antigenic epitopes may affect significantly the presentation of these epitopes in the complex with MHC-1 (Yellen-Shaw et al., 1997). Therefore, an important stage in the design of artificial multiepitope vaccines is calculation of the flanking regions between epitopes in order to ensure the efficient generation of these epitopes during processing of the vaccine antigen. This problem can be solved with the use of an index of the efficiency of epitope generation that we are proposing.

In addition, this approach can be applied to prediction of antigenic determinants and evaluation of algorithms for prediction of cleavage sites.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); US Department of Energy (grant № 535228 CFDA 81.049), and Civil Research and Development Foundation (grant № RB0-1276).

References

1. Altuvia Y., Margalit H. (2000). Sequence signals for generation of antigenic peptides by the proteasome: implications for proteasomal cleavage mechanism. J. Mol. Biol. 295: 879–890.
2. Ciechanover A. (1998). The ubiquitin-proteasome pathway: on protein death and cell life. EMBO J. 17: 7151–7160.
3. Holzhutter H.G., Frommel C., Kloetzel P.M. (1999). A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20S proteasome. J. Mol. Biol. 286: 1251–1265.
4. Kesmir C., Nussbaum A., Schild H., Detours V., Brunak S. (2002). Prediction of proteasome cleavage motifs by neural networks. Prot. Eng. 15: 287–296.
5. Lauvau G., Kakimi K., Niedermann G., Ostankovitch M., Yotnda P., Firat H., Chisari F.V., van Endert P.M. (1999). Human transporters associated with antigen processing (TAPs) select epitope precursor peptides for processing in the endoplasmatic reticulum and presentation to T cells. J. Exp. Med. 190: 1227–1240.
6. Nikolaev S.V., Afonnikov D.A., Ivanisenko V.A., Bazhan S.I., Kolchanov N.A. (2002). Comparison of methods for predicting proteasome cleavage motifs. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
7. Orlowski M., Wilk S. (2000). Catalytic activities of the 20S proteasome, a multicatalitic proteinase complex. Arch. Biochem. Biophys. 383: 1–16.
8. Rammensee H.-G., Bachmann J., Emmerich N.N., Bachor O.A., Stevanovic S. (1999). SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics. 50: 213–219.

9. Yellen-Shaw A.J., Wherry E.J., Dubois G.C., Eisenlohr L.C. (1997). Point mutation flanking a CTL epitope ablates *in vitro* and *in vivo* recognition of a full-length viral protein. J. Immunol. 158: 3227–3234.

BENCHMARKING OF PROGRAMS FOR RECOGNITION OF TRANSMEMBRANE SEGMENTS IN TRANSPORTER PROTEINS

[1] Sadovskaya N.S., [2] Sutormin R.A., [3] Rakhmaninova A.B., [2*] Gelfand M.S.

[1] Institute of Information Transmission Problems, RAS, Moscow, Russia
[2] GosNIIGenetika, Moscow, Russia
[3] Integrated Genomics, Moscow, Russia
e-mail: misha@imb.imb.ac.ru
[*]Corresponding author

Key words: *secondary transporter, bacteria, transmembrane segment, prediction, benchmarking*

Introduction

Mapping of transmembrane (TM) helices in integral membrane proteins, e.g. transporters, is an important area of bioinformatics. Most algorithms predict not only individual TM-helices, but the topology of the protein as well, which is interesting from both functional and evolutionary points of view. Due to experimental differences, only about twenty structures of TM-proteins were solved by crystallographic analysis, and thus straightforward benchmarking of the prediction algorithms is impossible. We benchmark four most widely used TM-helix prediction servers using the consistency criterion: predictions for homologous proteins should be similar.

Methods

We make the following assumptions:

in a group of orthologous proteins the number and lengths of TM-segments should coincide;

when these proteins are aligned, the positions of the TM-segments also should coincide, that is, corresponding TM-segments in all proteins should map to one segment of the alignment.

Thus for a pair of aligned TM-proteins, the overlap value *QQ* (equal to the ratio of the intersection length of the predicted TM-segments to the union length) should be close to 1 (Fig.). To estimate the level of coincidence on the level of whole segments, we computed the *KFS* value defined as follows. Consider a pair of intersecting TM-segments *i*, *j* in two aligned proteins. Let kfs_{ij} be the ratio of the intersection length to the length of segment *i* (l_i). If thus computed $kfs_{i\,j} > 0.5$, re-set kfs_{ij} =1, otherwise, re-set kfs_{ij} =0. Similarly, compute kfs_{ji} for segment *j*. Finally, compute *KFS* as the sum of kfs_{ij} and kfs_{ji} over all segments *i* from the first protein and *j* from the second protein divided by the total number of TM-segments in both proteins. Again, *KFS* =1 for absolutely consistent predictions. The use of this value allows one to ignore small differences in mapping TM-segment boundaries.

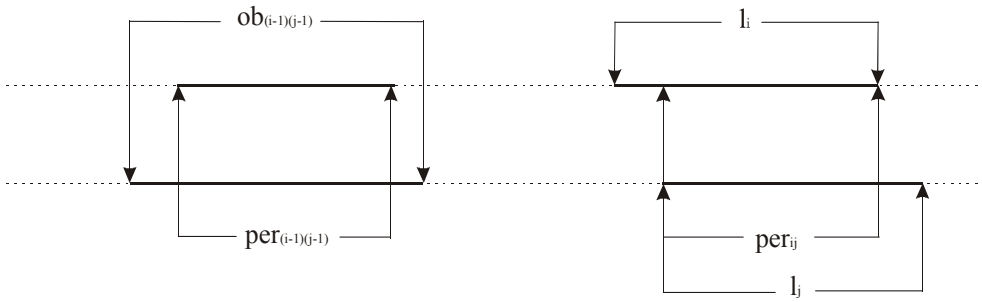


Fig. Computing QQ and KFS. Solid line: TM-segments, broken line: loops of aligned orthologous TM-proteins. $QQ = \Sigma_{i,j}\, per_{ij}/\Sigma_{i,j}\, ob_{ij}$, $kfs_{ij} = per/l_i$, $kfs_{ij} = per/l_i$, $KFS = \Sigma_{i,j}\,(kfs_{ij} + kfs_{ij}) / (tm_1 + tm_2)$, where ob_{ij} is the union of TM-segments, per_{ij} is the intersection of TM-segments in the two proteins, l_i and l_j are the lengths, tm_1 and tm_2 are the numbers of TM-segments in the two proteins.

One representative was selected from each of the following bacterial secondary transporter families (class TC.2A in the Saier classification [1, 2]): GNTP (2A.8), MHS (2A.1.6), NCS2 (2A.40). For each protein three groups of orthologs were selected from the ERGO database [3], corresponding to the identity levels 40-60%, 60-80%, and 80-100%. Total 26 proteins were analyzed (Table 1).

The obtained sequences were aligned with ClustalW [4]. TM-segments were predicted by four servers:

DAS (http://www.sbc.su.se/~miklos/DAS/) [5],

TMAP (http://www.mbb.ki.se/tmap/) [6],

TMHMM1.0 (http://www.cbs.dtu.dk/services/TMHMM-2.0) [7],

TMPred (http://www.ch.embnet.org/software/TMPRED_form.html).

Overlapping segments predicted by a server in one protein were merged. All pairs of orthologs were used to compute the consistency indices *QQ* and *KFS*.

Table 1. The number of proteins in each group.

Family \ identity	40-60%	60-80%	80-100%
MHS	3	5	4
NCS2	8	0	4
GNTP	13	1	3

Results and Discussion

The average values of the consistency indices for each family and each range of pairwise identity are given in Table 2. The values for two best servers, TMHMM and TMPRED are comparable. It should be noted, however, that these data are preliminary. We are currently performing a larger-scale study involving all available bacterial secondary transporters and more servers.

Table 2. Consistency values (see the text for explanation).

		DAS	DAS	TMAP	TMAP	TMHMM	TMHMM	TMPRED	TMPRED
		qq	kfs	qq	kfs	qq	kfs	qq	kfs
MHS	40-60%	0,259	0,345	0,574	0,774	0,581	0,696	0,489	0,591
	60-80%	0.552	0,770	0,640	0,881	0,715	0,969	0,638	0,883
	80-100%	0,886	0,975	0,946	1,000	0,966	1,000	0,902	0,994
NCS2	40-60%	0,362	0,499	0,525	0,794	0,562	0,784	0,502	0,741
	60-80%	0,815	0,947	0,613	0,823	0,858	0,976	0,802	1,000
	80-100%	0,925	0,980	0,837	0,918	0,946	0,981	0,925	0,981
GNTP	40-60%	0,365	0,499	0,564	0,784	0,551	0,735	0,482	0,671
	60-80%	0,469	0,692	0,619	0,866	0,579	0,841	0,567	0,855
	80-100%	0,827	0,945	0,769	0,916	0,778	0,906	0,828	0,935

Acknowledgments

We are grateful to A.A.Mironov, V.Yu.Makeev, and O.V.Kalinina for useful discussions. This study was partially supported by grants from RFBR (00-15-99362), INTAS (99-1476), HHMI (55000309) and LICR (CRDF RB0-1268).

References

1. Saier M.H.Jr. 1999. A functional-phylogenetic system for the classification of transport proteins. Cell Biochem. 84-94.
2. Saier M.H.Jr. 2000. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol. Mol. Biol. Rev. 64(2):354-411.
3. Overbeek R., Larsen N., Pusch G.D., D'Souza M., Selkov E.Jr, Kyrpides N., Fonstein M., Maltsev N., Selkov E. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucl. Acids Res. 28:123-125.
4. Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Acids Res. 22:4673-4680.
5. Cserzo M., Wallin E., Simon I., von Heijne G., Elofsson A. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng. 10:673-676.
6. Persson B., Argos P. 1996. Topology prediction of membrane proteins. Protein Sci. 5:363-371.
7. Sonnhammer E.L., von Heijne G., Krogh A. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. 6:175-182.

PROTEIN PROFILES BASED ON STRUCTURAL DESCRIPTORS OF AMINO ACID RESIDUES

**** Sobolev B.N., Fomenko A.E., Filimonov D.A., Poroikov V.V.***

V.N.Orekhovich Institute of Biomedical Chemistry, RAMS, Moscow, Russia, e-mail: boris@ibmh.msk.su
*Corresponding author: boris@ibmh.msk.su

Key words: *protein family, alignment, profile, structural descriptors, computer analysis*

Resume

Motivation: The protein family profiles derived from sequence alignments allow to reveal common functional features of annotated proteins. The problem of more detailed characterization is far from complete solution. In this reason, new approaches to provide more sensitive recognition of functional regions have to be developed.

Results: Earlier we developed structural descriptors (MNA-descriptors), successfully used for structural comparison of low-molecular compounds. In this study, we applied MNA-descriptors to describe amino acid residues separately or within alignment columns. Profiles based on MNA-descriptors were calculated for structurally aligned proteins related to the trypsin family. Two respective score types were selected from seven ones in preliminary studies on the small sequence set, and then the most conservative profile fragments were separately scanned SWISSPROT database. We had shown that MNA-descriptors could be used to found the conservative functional regions.

Introduction

The existent methods designed to calculate the amino acid sequence profiles use amino acid occurrence frequencies in the separate alignment columns (see, e.g.: Gribskov, 1994; Bateman et al, 2002; Jones, Swindells, 2002). This approach increases the sensitivity and reliability in recognition the common features of amino acid sequences belonged to the protein family. However, establishing of the remote homology and more detailed characterization of novel proteins (e.g., substrate specificity prediction) are solved ambiguously. The task is complicated due to the large sequence diversity in the remote homologues. We propose to apply new molecular structural descriptors, called MNA (Multilevel Neighborhoods of Atoms), to characterize the columns of protein alignment. In this approach, each aligned position is described by the set of structural descriptors calculated for the each amino acid residue of the alignment column. We suggest that profiles based on molecular descriptors help to reveal the molecular fragments that are common for profile positions and scanned sequence residues. MNA-descriptors developed in our laboratory were successfully employed earlier for prediction the drug activity of low-molecular compounds (Filimonov et al., 1999). The first results show that approach based on MNA-descriptors could be applied for protein profile calculation.

Methods and Algorithms

MNA-descriptors are calculated from the structural formula of the compound. These descriptors take into account the neighboring atoms in environment of the described atoms. Per se, separate descriptor presents the molecular fragment. In our study we created list of unique MNA-descriptors for each amino acid residue. Example of MNA descriptors calculated for Alanine is presented on Fig. 1. Amino acid residues were described by the different sets of descriptors that ranged from nine (for Glycine) to the twenty-eight (for Tryptophan). We used one hundred and thirty three unique descriptors to describe twenty amino acids. These descriptors were numbered that ensure to transform descriptor sets to sets of integer values. Three of the descriptors were common for all amino acid residues. Thus, they were excluded from the list.

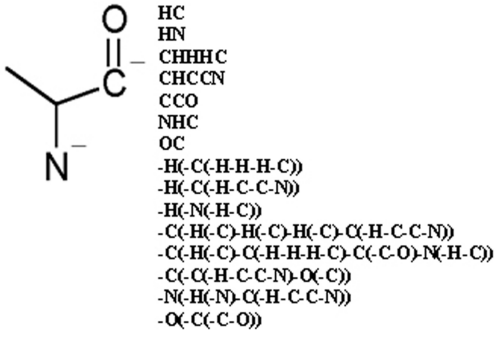


Fig. 1. MNA-descriptors for Alanine.

The alignment of proteins was obtained from multiple structure superposition (3D-alignment) performed by the CE algorithm (Shindyalov, Bourne, 1998) (http://cl/sdsc.edu/ce.html).

Each column in alignment was described by intersection and union of MNA-descriptor sets (Fig. 2). Thus, the each profile position is formed by two sets presented intersection and union of descriptors respectively. We test seven different scoring procedures on the small set of amino acid sequences, including the proteins related and non-related to the trypsin family:

Q include C (score 1)

N(Q and C) / N(Q) (score 2)

N(Q and U) / N(Q) (score 3)

N(Q and C) / N(C) (score 4)

N(Q and U) / N(U) (score 5)

N(Q and C) / N(Q or C) (score 6)

N(Q and U) / N(Q or U), (score 7),

where Q, C and U are descriptor sets of separate residue (Q), profile position intersection (C) and profile position union (U). N(X) is the number of descriptors in the X set.

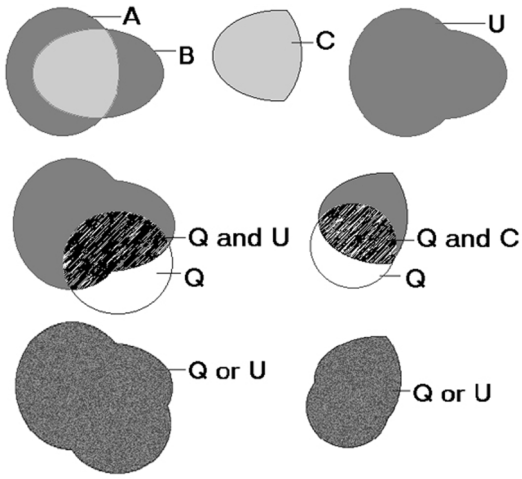


Fig. 2. Venn's diagrams for operations on MNA-descriptor sets. A, B – MNA-descriptor sets of amino acid residues in alignment column; C, U – intersection and union of the MNA-descriptor sets respectively; Q – MNA-descriptor set of amino acid residue of scanned sequences.

We tested all the mentioned procedures for matching the corresponding regions with maximal score values as well as discriminating the trypsin and non-trypsin proteins. The best results were defined for procedures 3 and 7. Based on those procedures we introduced S1 and S2 scores calculated as the maximal values normalized to average ones for each sequence:

S1 = max (score 3) / average (score 3),

S2 = max (score 7) / average (score 7),

The stand-alone version of SWISSPROT database was used for profile scanning. Records related to protein structures used for 3D alignment were excluded from the scanning. The proteins related to the trypsin family are detected by links to PFAM database (PF0089). Below, eukaryotic trypsin proteins, non-eukaryotic trypsin proteins and non-trypsin ones are designated as *eutry*, *non-eutry* and *non-try,* respectively. The thirty-seven 3D structures of trypsin proteins were selected for constructing the profiles.

Results and Discussion

The SWISSPROT database was scanned with two profiles. The first one was constructed from both eukaryotic and prokaryotic structures (profile I), and the second one – from eukaryotic structures only (profile II). We had chosen three conservative fragments in the profiles, which contain active site residues, and scanned the SWISSPROT database. All profile fragments revealed the clear difference between the *eutry* and *non-eutry* proteins. The most significant results were obtained for His-containing fragment of the profile II (see the Fig. 3). The number of scanned sequences that fell in the respective S1 and S2 range was normalized to general number of the respective protein category (*eutry*, *non-eutry* and *non-try*). Testing profile II at His-containing fragment also revealed the separation between the trypsin and non-trypsin protein, though it was not so significant. Two other fragments, containing catalytic Ser and Asp, respectively, also showed difference that more stretched in case of eukaryotic profile (these results are not displayed).

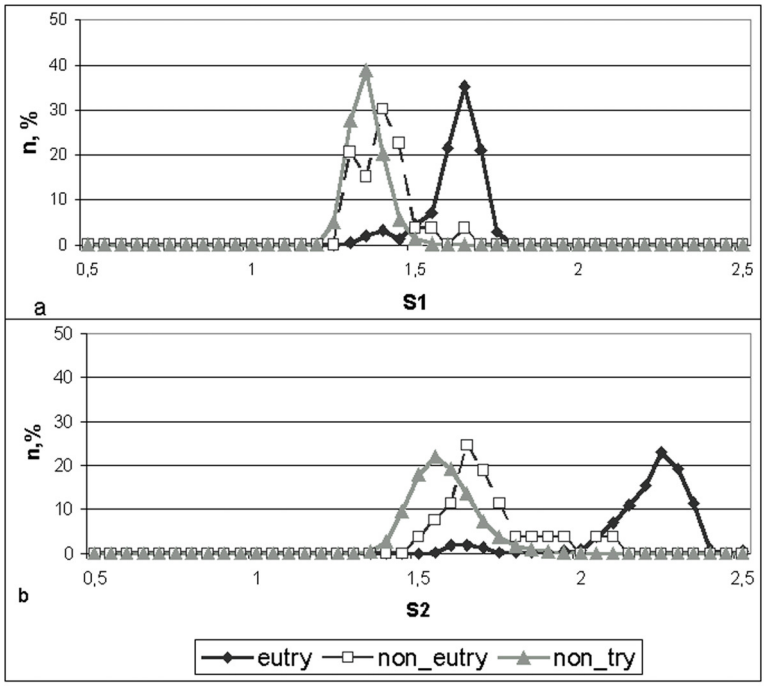


Fig. 3. Result of SWISSPROT scanning with profile fragment, containing catalytic His. n,% - frequencies of proteins (in percentage) with respective score value (S1 or S2).

Our results revealed that MNA-descriptors could be applied for constructing the profiles and database searching to annotate of the proteins. Relatively number of high scores was obtained for non-trypsin proteins that can be explained by using the short conservative fragments in the separate mode. At this point, combined scoring along all the contiguous regions have to increase discriminating power.

Acknowledgements

The work was supported by Russian Foundation for Basic Research (grants № 01-04-48710 and 02-04-6619).

References

1. Gribskov M. (1994) Profile analysis. Methods Mol. Biol. 25, 247-266.
2. Bateman A., Birney E., Cerruti L., Durbin R., Etwiller L., Eddy S.R., Griffiths-Jones S., Howe K.L., Marshall M., Sonnhammer E.L. (2002) The Pfam protein families database. Nucl. Acids Res. 30, 276-280.
3. Jones D.T., Swindells M.D. (2002) Getting the most from PSI-BLAST. Trends Biochem Sci. 27, 161-164.
4. Filimonov D., Poroikov V., Borodina Yu., Gloriozova T. (1999) Chemical Similarity Assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. J. Chem. Inf. Comput. Sci. 39 (4), 666-670.
5. Shindyalov I.N., Bourne P.T. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng. 11, 739-747.

COMPUTATIONAL ANALYSIS OF POTENTIAL DISULPHIDE BRIDGES IN PLANT DNA TOPOISOMERASE I

[1]* Konstantinov Y.M., [2] Rogozin I.B., [2] Tarasenko V.I.

[1]Siberian Institute of Plant Physiology and Biochemistry, SB RAS, Irkutsk, 664033, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
*Corresponding author; e-mail: yukon@sifibr.irk.ru

Key words: *DNA topoisomerase I, gene, core domain, conservative cysteine, redox modulation, evolution*

Resume

DNA topoisomerases are the key enzymes of genetic information processing; the cell deprived of topoisomerases fails to make up for their absence and, thus, perishes. These enzymes participate in many fundamental genetic processes associated with separation of DNA strands such as replication, transcription, recombination and repair. Topoisomerases may also act as DNA strand transferases and catalyze recombination and transposition reactions (Pommier, 1998). Type I DNA topoisomerase (topo I) acts by making a transient nick on a single-strand of duplex DNA, passing another strand through the nick, and changing the linking number by one unit (Gupta et al., 1995). Beyond their normal physiological functions, eukaryotic topo I have been identified as the primary cellular target for a variety of antitumor agents (Pommier, 1998).

Plant cells contain topoisomerases not only in nucleus, but also in mitochondria and chloroplasts. No detailed molecular biological studies of plant topoisomerases I of nuclear and mitochondrial localization have been accomplished up to now - neither at the level of encoding genes, nor at the level of their products. It should be noted that molecular mechanisms of regulation of activity of this enzyme remain practically unexplored. These mechanisms apparently differ for genetic systems functioning in nucleus and organelles.

Based on the analysis of sequence and functional mapping of eukaryotic topoisomerase I, it is divided into four principal domains: a poorly conserved N-terminal domain, a highly conserved core domain, a poorly conserved linker domain, and a highly conserved C-terminal domain, including the active tyrosine site. Taking into account the data acquired earlier on the importance of sulfhydryl groups for the activity of DNA topoisomerases of plant origin, we have hypothesized that the protein molecule of this enzyme may contain redox sensitive regulatory cysteines ensuring sensitivity of topo I to the change of redox state of glutathione and, possibly, other low molecular biothiols in a cell. Experimental study of redox conditions' impact on the activity of DNA topoisomerase I in plants has revealed that under oxidative conditions created by addition of potassium ferricyanide or oxidized glutathione (GSSG), significant decrease of topoisomerase activity was observed, whereas under reducing conditions created by addition of sodium dithionite or reduced glutathione (GSH), an activation of the enzyme takes place. GSSG significantly exceeded potassium ferricyanide in the degree of inhibition of DNA relaxation activity of topoisomerase, at the same time GSH and sodium dithionite produced approximately the same activating effect. Impact of redox system of glutathione on topoisomerase activity observed in the course of experiments, in our view, indicates to possible participation of reactions of thiol-disulfide exchange at the level of cysteine residues within highly conservative core domain of topo I in its activity modulation.

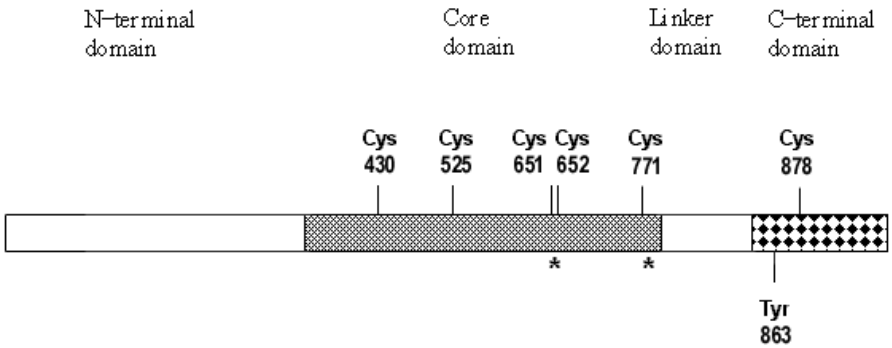


Fig. Domain structure of DNA topoisomerase I of carrot. The scheme is developed on the basis of the data of Balestrazzi et al. (1996). Cys – cysteine, Tyr – tyrosine of topoisomerase I active center. The numbers indicate the localization of highly conserved cysteine residues. Asterisks mark the sites, whose cysteines may form disulfide bond according to the 3D computer modeling data (Konstantinov et al., 2001).

The aim of the present work was to analyze the disulfide connectivity in DNA topoisomerase I of higher plants in order to reveal candidate cysteine residues. We have used a program developed by Fariselli and Casadio (2001). In this approach, the problem of predicting the disulfide connectivity in proteins is equated to a problem of finding the graph matching with the maximum weight. The graph vertices are the residues of cysteine-forming disulfide bridges, and the weight edges are contact potentials. Analysis of various control sets suggested that the method could be applicable to locate putative disulfide bridges in proteins (Fariselli, Casadio, 2001). Results of the carrot topo I analysis are shown in Table. Several potential disulfide bridges were found, earlier suggested candidate pair Cys 652 – Cys 771 (Konstantinov et al., 2001) was among 6 top-ranking candidates (Table). However, a pair Cys 430 – Cys 525 has a significantly higher potential to form the disulfide bridge. Suggested candidate disulfide bridges need further experimental investigation.

Table. A Fariselli-Casadio potential of the disulfide connectivity.

#1	#2	Potential
Cys 31	Cys 250	0.048
Cys 31	Cys 430	0.089
Cys 31	Cys 525	0.189
Cys 31	Cys 651	0.243
Cys 31	Cys 652	**0.372**
Cys 31	Cys 713	0.213
Cys 31	Cys 771	0.028
Cys 250	Cys 430	0.044
Cys 250	Cys 525	0.050
Cys 250	Cys 651	0.139
Cys 250	Cys 652	0.154
Cys 250	Cys 713	0.064
Cys 250	Cys 771	0.036
Cys 250	Cys 878	0.167
Cys 430	Cys 525	**0.678**
Cys 430	Cys 651	0.274
Cys 430	Cys 652	0.171
Cys 430	Cys 713	0.016
Cys 430	Cys 771	0.094
Cys 430	Cys 878	0.032
Cys 525	Cys 651	0.136
Cys 525	Cys 652	0.101
Cys 525	Cys 713	0.043
Cys 525	Cys 771	0.009
Cys 525	Cys 878	0.153
Cys 651	Cys 652	0.042
Cys 651	Cys 713	**0.367**
Cys 651	Cys 771	0.196
Cys 651	Cys 878	0.054
Cys 652	Cys 713	0.213
Cys 652	Cys 771	**0.302**
Cys 652	Cys 878	0.066
Cys 713	Cys 771	**0.673**
Cys 713	Cys 878	0.040
Cys 771	Cys 878	**0.448**

Acknowledgement

The study is supported by the Russian Foundation of Basic Research (grant № 01-04-48162).

References

1.Balestarzzi A., Toscano I., Bernacchia G., Luo M., Otte S., Carbonera D. (1996) Cloning of a cDNA encoding DNA topoisomerase I in *Daucus carota* and expression analysis in relation to cell proliferation. Gene. 183: 183-190.

2. Fariselli P., Casadio R. (2001) Prediction of disulfide connectivity in proteins. Bioinformatics. 17: 957-964.

3. Gupta M., Fujimori A., Pommier Y. (1995) Eukaryotic DNA topoisomerases I. Biochim. Biophys. Acta. 1262: 1-14.

4. Konstantinov Y.M., Tarasenko V.I., Rogozin I.B. (2001) Redox modulation of activity of mitochondrial topoisomerase I from carrot (*Daucus carota*). Dokl. Akad. Nauk (Mosk.). 377: 263-265.

5. Pommier Y. (1998) Diversity of DNA topoisomerases I and inhibitors. Biochimie. 80: 255-270.

LOGICAL ANALYSIS OF DATA APPROACH TO THE PREDICTION OF PROTEIN SECONDARY STRUCTURES

[1] Błażewicz J., [2] Hammer P.L., *[1] Łukasiak P.

[1] Institute of Computing Sciences, Poznan University of Technology, Poznan, POLAND
Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, POLAND
[2] Rutgers Center for Operations Research, Rutgers University, New Jersey, USA
e-mail: Piotr.Lukasiak@cs.put.poznan.pl
*Corresponding author

Key words: *logical analysis of data, protein prediction, protein secondary structure, machine learning*

Resume

Motivation: The reason that this problem is so important is that the structure of protein is directly dependent on its function. Experimental structure determination, or structure prediction, aids the elucidation of protein function; conversely, synthetic protein sequences might be designed so that the protein performs a desired function. The study of protein structure is therefore not only of fundamental scientific interest in terms of understanding biochemical processes, but also produces very valuable practical benefits.

Results: The obtained results over 70% for three classes of secondary structures are similar or better as compared with other methods for the protein prediction. A comparison has been made with the PHD algorithm and algorithm based on the Rough Set theory. During experiment the set of the most promising amino acids properties has been extracted for secondary structure description. LAD generated simple and strong rules which could be easily interpreted by biologists

Availability: available on request from the authors: Piotr.Lukasiak@cs.put.poznan.pl

Introduction

The first level of the protein structure, termed primary structure, refers just to the sequence of amino acids in the protein. Decades ago it was found that polypeptide chains can sometimes fold into regular structures; that is, structures which are the same in shape for different polypeptides. These structures create the second level of protein structure. When one looks at an actual polypeptide chain, its final shape is made up of secondary structures, perhaps super-secondary structural features, and some apparently random conformations. This overall structure is referred to as the tertiary structure. The three-dimensional structure of proteins is uniquely determined by its primary structure.

The widely used standard sequence search techniques like BLAST, FASTA searches of sequence databases have very good accuracy when used with care. The most widely used methods are currently the statistics-based GOR method, for its algorithmic simplicity and easy implementation, and the PHD program of Rost and Sander (Rost, Sander, 1993; Rost, 2000). The SSPAL method of Salamov and Solovyev uses multiple overlapping local alignments with sequences of known secondary structure in a nearest neighbour-like way and achieves 71% accuracy, which is uncharacteristic of a single sequence method (their multiple sequence nearest neighbour method NNSSP (Salamov, Solovyev, 1997) is 72% accurate by the same measures). The other popular solution is Monte Carlo method (Skolnick, Kolinski, 1999) trying to determine the structure which minimizes free energy. From the above overview it follows that such tools as machine learning are still needed because it is often difficult for humans to perceive patterns in data, even though strong patterns exist. The idea to create a tool helping molecular biologists was the main reason to choose the new rule-based method – Logical Analysis of Data (Boros et al., 1996).

The Method

The goal of the analysis described in this paper is to create a system which allows to receive as the output the protein secondary structure, based on its primary structure being an input, and to find rules responsible for this effect.

Let $W = \{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,Y,V,W\}$ be a set of all amino acids where each letter corresponds to a different amino acid.

The word s is called a protein primary structure on the condition that letters in this word are in the same order as amino acids in the protein chain are. Let the length of the word s be denoted as $C(s)$ and $A(s,j)$ denote an element of word s, where j is an integer number from the set $[1,C(p)]$.

In a similar way, a representation for the secondary structures as for the primary ones, can be created. Let $F = \{H,E,X\}$ represents a set of all secondary structures, where each letter corresponds to a different secondary structure. A secondary structure is represented here by a word on the relevant alphabet of secondary structures: each kind of a secondary structure has its own unique letter. Let us denote this word by d, where the length of word d is equal to the length of word s.

Now, we may define the problem as the one consisting in finding a secondary structure of a protein (in a form of word *d*), based on the protein primary structure (i.e. word *s*). Moreover, for each element *A(s,j)* one should assign an element *A(d,j)* in the way that the obtained secondary structure *r* is as close as possible to a real secondary structure of the considered protein.

In the paper, the Logical Analysis of Data (LAD) algorithm is used for the above problem. Examples were obtained from the Dictionary of Secondary Structures of Proteins (DSSP). DSSP contains a description of secondary structures for entries from the Brookhaven Protein Data Base.

The following three sets of secondary structures have been created for the experiments: helix (H) consisting of: α-helix (structure denoted by H in DSSP), 3_{10}-helix (G) and π-helix (I); β-strand (E) consisting of E structure in DSSP; the rest (X) consisting of structures belonging neither to set H nor to set E.

Because of a complexity of the algorithm of Logical Analysis of Data it is hard to present all aspects of this method. An interested reader is referred to (Boros et al., 1996; Boros et al., 1997) for a more detailed description of the Logical Analysis of Data method.

The first step one has to do, is to prepare a set of observations (based on a protein sequence) to be acceptable by the LAD. Below an example is presented, that illustrates the way a protein chain is changed into a set of observations. Let us consider a protein chain called *4gr1* (in PDB). The first and the last fifteen amino acids in the sequence are shown below:

VASYDYLVIGGGSGG ... VAIHPTSSEELVTLR

For every amino acid the corresponding secondary structure in DSSP is given as follows:

_EE_SEEEE__SHHH ... ___SS_SGGGGGS__

One may change this structure into secondary structures involving three main secondary structures only in the way depicted below:

XEEXXEEEEXXXHHH ... XXXXXXXHHHHHXXX

At the end of a chain consisting of n amino acids one obtains a set consisting of n observations as shown in Table 1.

A window of length 6 generates an observation with 6 attributes (a_{-3}, a_{-2}, a_{-1}, a_0, a_{+1}, a_{+2}) representing a secondary structure corresponding to the amino acid located in place a_0. Of course, at this moment all values of attributes are symbols of amino acids. Secondary structures on the boundaries have been omitted from the consideration.

Table 1. An example transformation from a sequence to a set of observations.

#	Condition attributes $a_{-3}a_{-2}a_{-1}a_0a_{+1}a_{+2}$	Code in DSSP	Codes of the three secondary structure
1	* * V A S Y	E	E
2	* V A S Y D		X
3	V A S Y D Y	S	X

The last step of the preprocessing is to replace in each observation symbols of amino acids (treated as attributes) with numbers representing relevant properties of amino acids. During experiment only the physical and chemical properties of the amino acids offered by ProtScale have been taken into account. Originally we considered 54 properties, but, after a discussion with domain experts, 28 of them have been chosen for the first experiments. For a detailed description of all properties see (Blazewicz et al, 2001). At the end from the set of 54 properties, 6 of them have been extracted which had the most important influence on the created secondary structures.

Results and Discussion

85 protein chains have been chosen into consideration. Using FASTA algorithm we checked annotated alignment of related sequences in the considered set. FASTA recognized 50 sequences with no alignment, 9 groups consisted of 2 aligning sequences, 2 groups consisted of 3 aligning sequences and one group consisted of 11 aligning sequences. Based on these 85 protein chains about 20 000 observations have been created using the algorithm described. As one mentioned above, originally we considered 54 properties, but after a discussion with domain experts 28 of them have been chosen for the experiment.

Hydrophobicity scale (pi-r) and *molecular weight* were the best properties for class H, *average surrounding hydrophobicity* and *bulkiness* gave the best results for class E and *polarity* and *hydropathicity* were the best ones for class X .

For the next part of experiments one decided to create a set consisting of 2000 observations and apply 5-fold cross validation test. These 2000 observations were selected randomly from the set of 20000 observations described above. The number of observations were smaller because we decided to enlarge the number of attributes. Now, each observation consisted of 12 attributes, where the first 6 corresponded to one property, and the last six corresponded to another property.

The mix of the properties increased the accuracy by 10% as compared to the results obtained using each property separately (Blazewicz et al, 2001).

In Figure one can see results obtained using LAD with results obtained using PHD method. For PHD only results for class H and E were shown, because only the results for these two structures were presented in (Rost, Sander, 1993). Average results are similar and none of the methods proved its superiority, but LAD gave also rules which could explain for biologists some features of the phenomenon like protein prediction problem. Some rules explored during experiments are presented in Table 2.

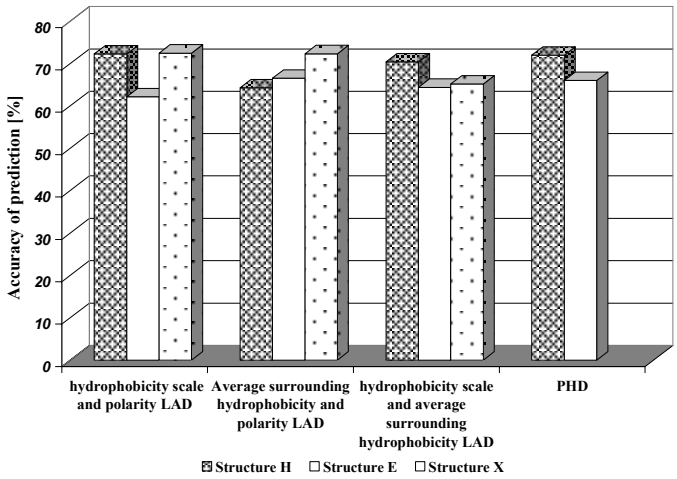


Fig. Comparison of prediction between LAD and PHD

Table 2. Example of rules for class H.

#	a_{-3}	a_{-2}	a_{-1}	a_0	a_{+1}	a_{+2}	Property
1	>-0,705	———	>0,285	<0,065	———	<-0,620	Hydrophobicity scale (pi-r)
2	<-0,620	<-0,130	———	>1,795	———	>-0,020	
3	———	>1,745	<0,195	>1,225	>1,795	———	

In the context of machine learning algorithms LAD gave results similar and better to the best standing alone methods used for protein prediction problem. In the molecular biology context LAD generates simple and strong rules which could be easily interpreted by biologists. It has been shown above that not only certain global features, such as the presence of helix or strand structure can be predicted with a usable accuracy and reliability using the method like LAD but also the description of the secondary structure of each residue can be predicted.

References

1. Blazewicz J., Hammer P.L., Lukasiak P. (2001) Prediction of protein secondary structure using Logical Analysis of Data algorithm. Computational Methods in Science and Technology. 7(1).
2. Boros E., Hammer P.L., Ibaraki T., Kogan A., Mayoraz E., Muchnik I. (1996) An implementation of logical analysis of data. Rutcor Research Report. 22-96.
3. Boros E., Hammer P.L., Ibaraki T., Kogan A. (1997) Logical Analysis of Numerical Data. Rutcor Research Raport. 04-97.
4. King R.D., Sternberg M.J.E. (1990) Machine learning approach for the prediction of protein secondary structure. J. of Mol. Biol. 216, 441-457.
5. Rost B. (2000) PROF: predicting one-dimensional protein structure by profile based neural networks, unpublished.
6. Rost B., Sander C. (1993) Prediction of protein secondary structure at better than 70 % accuracy. J. of Mol. Biol. 232, 584-599.
7. Salamov A.A., SolovyevV.V. (1997) Protein secondary structure prediction using local alignments. J. of Mol. Biol. 268, , 31-36.
8. Skolnick J., Kolinski A. (1999) Monte Carlo approaches to the protein folding problem. In Ferguson D., Siepmann J.I., Truhlar D.G. (eds). Monte Carlo Methods in Chemical Physics, Advances in Chemical Physics, John Wiley & Sons. 105, 203-242.

NONSTANDARD APPROACH FOR α–HELICES ELUCIDATION

[1] Kilosanidze G.T., [2] Kutsenko A.S., [1] Esipova N.G., [1*] Tumanyan V.G.

[1] Engelhardt Institute of Molecular Biology, RAS, Moscow, 119991, Russia
[2] Centre for Genomics and Bioinformatics, Karolinska Institute, P.O. Box 280, S-171 77 Stockholm, Sweden
[*]The corresponding author. e-mail: tuman@imb.imb.ac.ru

Key words: *DNA-proteins, secondary structure, α-helices prediction, molecular mechanics*

Resume

Motivation: There is a necessity in understanding of physical grounds of secondary structure formation in globular proteins. All modern methods for secondary structure prediction have serious limitations originated from their statistical nature. Further progress in this field may be attained only on the basis of physical rules governing secondary structure formation. In this connection the forces stabilizing structure must be directly estimated by theoretical and/or experimental approaches. In the suggested method, the whole protein is considered as one α-helical segment for which molecular mechanics energy estimation is performed. One may expect more favorable energy at the regions where real α-helical segments occur. This will be in contrast with the regions that have non-helical conformation in the native protein.

Results: Molecular mechanics computations for more than 70 proteins have been fulfilled for model α-helical conformations. The mean accuracy of α-helices recognition is above 80% with equal number of false positive and false negative predictions. This nonstatistical approach permits to recover the interactions stabilizing α-helices and their dependence on the sequence of amino acid residues.

Availability: the method can be readily implemented by using standard molecular mechanics program.

Introduction

There is apparent progress in modern methods of secondary structure prediction. However, rather high prediction accuracy is attained on the basis of statistical approach or even direct resemblance of the segments at the level of primary structure (Baldi et al., 2000) in contrast to previous methods based on physical positions (Lim, 1974; Ptitsyn, Finkelstein, 1983). As a consequence it is unlikely to expect a good prediction for a protein which does not have a related protein with known spatial structure. Due to statistical or resemblance criteria fine physical grounds of secondary structure formation rest undiscovered. Therefore it is challenging to design a new physical approach for accurate secondary structure prediction. Any physical method with high prediction power would provide an insight into basic grounds of secondary structure folding.

Method

We developed principally new approach utilizing molecular mechanics calculations (Kilosanidze et al., 2002). A protein chain is folded in continuos α-helix by fixing ϕ and ψ angles at appropriate values. Optimal conformations of side chains are attained by standard Monte Carlo procedure as described in (Abagyan, Totrov, 1999). At the end of optimization procedure backbone (ϕ, ψ) angles are restrained with allowed 20° interval around ideal α-helix values. In order to account for constant and nearly constant contributions in energy varying from residue to residue we subtracted the energy of unrestrained conformation. This extended baseline conformation is characterized by values of (ϕ, ψ) angles near 180° as in (Pitera et al., 2000). Possible hindrances in the extended conformation are removed by additional optimization to adjust both (ϕ, ψ) angles in the main chain and χ-angles in the side chains. After this by subtracting the energy of baseline conformation from α-helical conformation for constituent oligopeptides (ordinary pentapetides), the resultant energy profile is constructed. Energy profile for extended baseline conformation is rather smooth, thus the profile for α–helical conformation often may be used alone. Profiles are constructing by energy estimation for overlapping fragments and the curves are practically unaffected by fragment length. Most of calculations were performs for pentapeptides.

Sample The sample included more than 70 proteins selected from PDB database by chance. Among them are proteins with PDB codes 1baz, 1bm9, 1igd, 10pc, 1pdn, 1pdv, 1ptf, 2chs, 1a92, 1aew, 1bbb, 1eq7 representing α, α+β and α/β classes.

Results

Modeling conformations were constructed by using the ICM program (Abagyan et al., 1994). Figure 1 illustrates the choice of best threshold value. The histogram includes all proteins under study. One can see that 26 kcal/mol value provides 80.09% level of accuracy for standard parameter $Q_{3\alpha}$ for two-state prediction (Shulz, Schirmer, 1979) with equivalence of false positive and false negative predictions. Figures 2 and 3represents recognition of α-helical patterns for 1a92 and 1aew, correspondingly.

There is close coincidence of minima on energy profiles for continuos model α-helix and experimental α-helical segments. Interhelical segments cannot adopt the proper α-helical structure for any side groups conformations. It is obvious from the figures that both of total energy and the van der Waals component minima on the profiles occupy the same positions. At the same time, van der Waals component is more convenient than electrostatic or hydrogen bonding since in their case, an estimated energy threshold can be universally used for α-helical fragment recognition in new proteins.

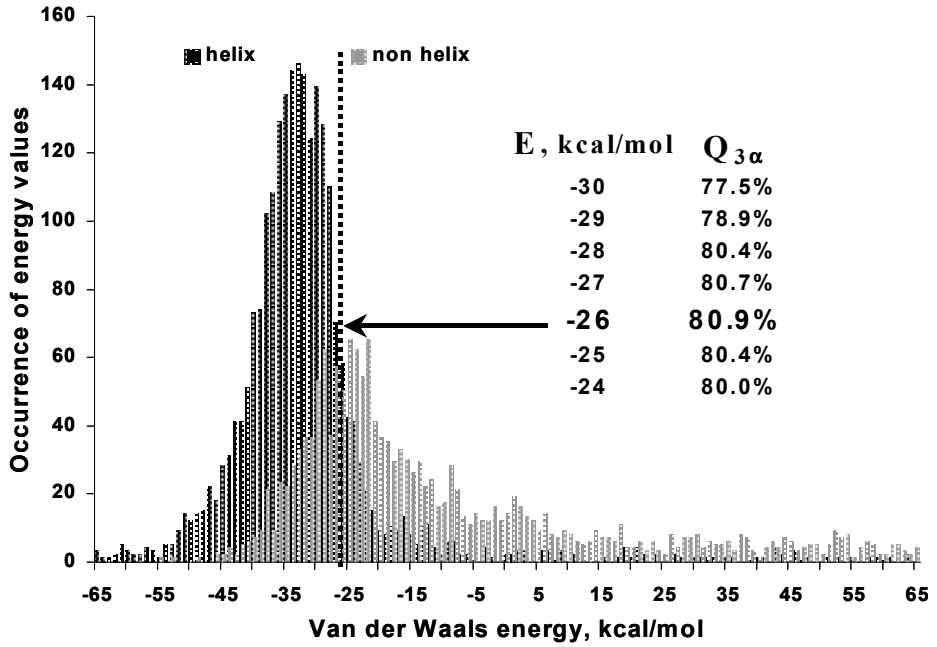


Fig. 1. Histogram for recognition of α-helices for the protein sample.

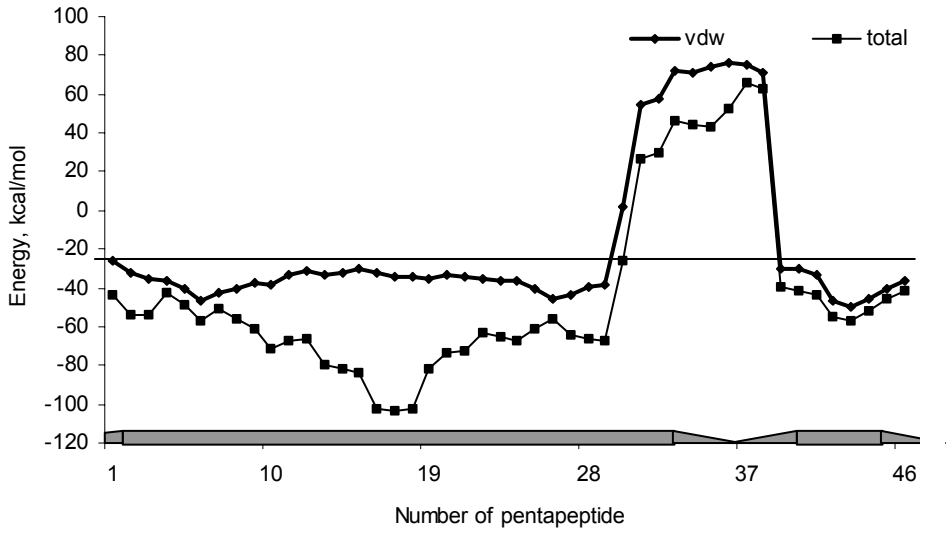


Fig. 2. Profiles of energy for oligomerization domain of hepatitis delta antigen (1a92).

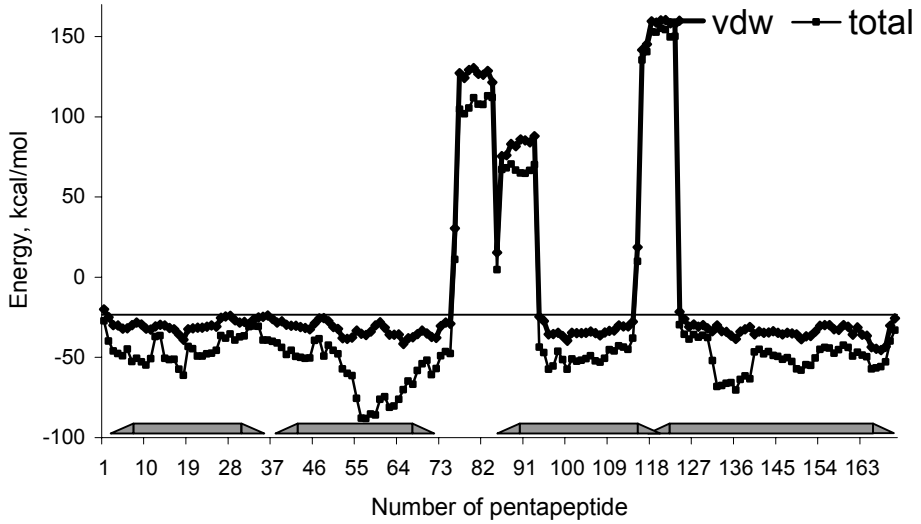


Fig. 3. Profiles of energy for L-chain horse apoferritin (1aew). Residues 6-175.

Conclusion

This nonstatistical approach allows one to observe real interactions stabilizing α-helices in connection to the sequence of amino acid residues. It is tempting to hope that the described method is promising for *a priori* secondary structure recognition, analysis and prediction.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (Grant 02-04-49114).

References

1. Baldi P., Brunak S., Chauvin Y., Andersen C.A., Nielsen H. (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics. 16, 412-424.
2. Lim V.I. (1974) Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. J. Mol. Biol. 88: 873-894.
3. Ptitsyn O.B., Finkelstein A.V. (1983) Theory of protein secondary structure and algorithm of its prediction. Biopolymers. 22: 15-25.
4. Kilosanidze G.T., Kutsenko A.S., Esipova N.G., Tumanyan V.G. (2002) Use of molecular mechanics for secondary structure prediction. Is it possible to reveal α-helix? FEBS Lett. 510: 13-16.
5. Abagyan R., Totrov M. (1999) Ab initio folding of peptides by the optimal-bias Monte Carlo minimization procedure. J. Comp. Phys. 151: 402-421.
6. Pitera J.W., Kollman P.A. (2000) Exhaustive mutagenesis in silico: multicoordinate free energy calculations on proteins and peptides. Proteins. 41: 385-397.
7. Abagyan R., Totrov M., Kuznetsov D. (1994) ICM: An efficient technique for structure predictions and design. J. Comp. Chem. 15: 488-506.
8. Schulz G.E., Schirmer R.H. Principles of Protein Structure. New York: Springer Verlag, 1979. 1-314.

SPIDER SILK FIBROUS PROTEIN β-STRUCTURE AND LARGE PERIODICAL PATTERNS

[1] Ragulina L.E., [2] Makeev V.Ju., [3] Esipova N.G., [3] Tumanyan V.G., [2] Bogush V.G., [2] Sidoruk K.S., [2] Debabov V.G.*

[1] Moscow Institute of Physics and Technology, 141700, Moscow, Russia
[2] State Scientific Center "GosNIIGenetika", 113545, Moscow, Russia
[3] Engelhardt Institute of Molecular Biology, RAS, 119991, Moscow, Russia
* The corresponding author. e-mail: lera_846@pisem.net

Key words: s*pider, web silk, protein, primary structure, periodical pattern, bioengineering*

Resume

Motivation: Spider web silk is a promising starting point for creation of a man-made protein fiber with exceptional properties. In order to keep the specific physical properties in engineered protein, one needs to retain its particular primary sequence structure with several systems of periodical patterns exhibited throughout the whole chain. This complicated periodical arrangement can be accurately identified with a symbolic Fourier transform accompanied with inter-specie comparison and then used to design a sequence of recombinant model protein. The type of this periodical pattern determines the protein secondary structure as well as the fibril packing. The structure suggested from the sequence features can be verified with the experimental results in order to choose the structure with optimal physicochemical properties out of the possible variants of bioengineered protein.

Results: For all types of spider web proteins contained in the databank we have identified characteristic periodical patterns with the help of the program SymFour. The distribution of periods was dissimilar from that characteristic for the collagen-like and α-helix-like types of silks and suggests a possible β-structure. Conservation of basic periodical structures was revealed by inter-specie comparison. For the longest fragment of spidroin II of *Nephila senegalensis* we assessed the distribution of the periodical patterns along the sequence and demonstrated several distinct sequence segments with their specific periodical patterns. We have designed a model amino acid sequence, which retains basic characteristic periodical patterns exhibited in the native spidroine II molecule, but which complies to the specific requirements applied by the practical procedure of recombinant protein engineering. By means of CD and FTIR we demonstrated that two conformations of polypeptide chain are simultaneously found in the solution, whereas the fibrous samples exhibit FTIR spectra characteristic for antiparallel β-structure.

Availability: SymFour program is available from makeev@imb.ac.ru on request.

Introduction

Spider silk has an outstanding physical properties, which makes this fiber a promising raw material for different technological applications (1990). These physical properties is a manifestation of a particular structure of a protein molecules the web fiber consists of (Madsen, 1999). The protein spatial structure at all its 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 charges distribution is associated with amino acid distribution along the sequence, which in fibrous – but not in globular proteins — display the periodic structure.

Methods and algorithms Each dragline web fiber contains two types of protein, spidroin I and spidroin II, with mass ratio ½. From all spidroin sequences contained in NCBI protein database (SwissProt+TrEmbl) those longer than 200 residues were extracted. There were found 17 sufficiently long fragments of spider silk proteins truncated at the N-end (7, sp I +10, sp II), of which one sequence, spidroin II of *Nephila senegalensis* was about 2000 residues, other were much shorter. All spidroin sequences (of both I and II types) contained a more conservative aperiodic C-end domain and a more variable semi-periodic domain of an unknown length. In our study we focused on the semi-periodic domain. A characteristic feature of this domain is poly-A stands of a variable length - from 4 to 8 residues - which were periodically located in the polypeptide sequence. When we aligned protein segments between these poly-A strands, it was found that for all species studied these segments contained one, two or three similar (repeated) units. The sequences of such units were made of a limited number of fundamental di- or tripeptides, such as QQ, GPG, GGY, etc. We observed that the number of units alternated within each segment between two poly-A strands. Thus, it was impossible to identify a repeated unit of any length, which could serve as a basis for protein engineering.

Alignments There as a number of fragments of spider silk proteins in the database, however almost all fragments are sequences of a limited length from the protein C-end (Gatesy et al., 2001). All these sequence are composed from the standard words and exhibit a periodical structure, however the "alphabet" of the standard words is specific for each particular specie. The lengths of the periods are also specie-specific and we failed to identify a common pattern

characteristic for spidroin protein of all species. However, in all cases the basic words contained poly-A, QQ, and GP combinations. The standard words are GPG, QQ, GY, but form common 10-letter words such as QQGPGGYGPG

```
QGPSGPGSAAAAAAA
GPGQQGPGGYGPGQQGPGGYGPGQQGPS          GPGSAAAAAAAAAA
GPGQQGPGGYGPGPQGPGGYGPGQQGPSGYGPGQQGPSGPGSAASAAAAA
GSGQQGPGGYGPGQQGPGGYGPGQQGPS          GPGSAAAAAAA
GPGQQGPGGYGPGQQGPGGYGPGQQGPS          GPGSAAAAAAAA
GPGQQGPGGYGPGQQGPGGYGPGQQGPS          GPGSAAAAAAA
GPGQQGPGGYGPGQQGPGGYGPGQQGPS          GPGSAAAAAAAAA
GPGQQGPGGYGPGQQGP     GQQGPS          GPGSAAAAAAA
GPGPQGPGGYGPGQQGPGGY      GPS          GPGSAAAAAAAA
GPGQQGPGGYGPGQQRPSGYGPGQQGPS          GPGSAAAAAAA
GPGQQGPGAY                GPS         GPGSAAAAA
     GLGGY           GPAQQGPS          GAGSAAAAAAA
     GPGGY           GPVQQGPS          GPGSAA
     GPGGY           GPAQQGPARY        GPGSAAAAAAAA
```

Fig. 1. A periodic alignment of the tail fragment of *Nephila madagascariensis* spidroin 2 protein.

Development of a model sequence. Spider silk analog protein engineering includes polymerization of some basic fragment. The final sequence can be obtained as a result of polymerization and merging of several initial monomers. Our experimental facilities does not allow us to use a monomer longer than 168 residues, we are limited with the number of participating monomers as well. Our main objective was to suggest an experimental design with the final sequence complying with the requirements above and retaining the basic system of periodical patterns of the native sequence, which we believe necessary for the molecular packing.

To this end we studied periodical patterns in the native protein sequence of *Nephilia madagascariensis* sp.II with the help of a symbolic Fourier transform algorithm.

Sequence Fourier analysis. The power spectrum of the whole 1953 amino acid long sequence was calculated and several outstanding peaks were identified, including

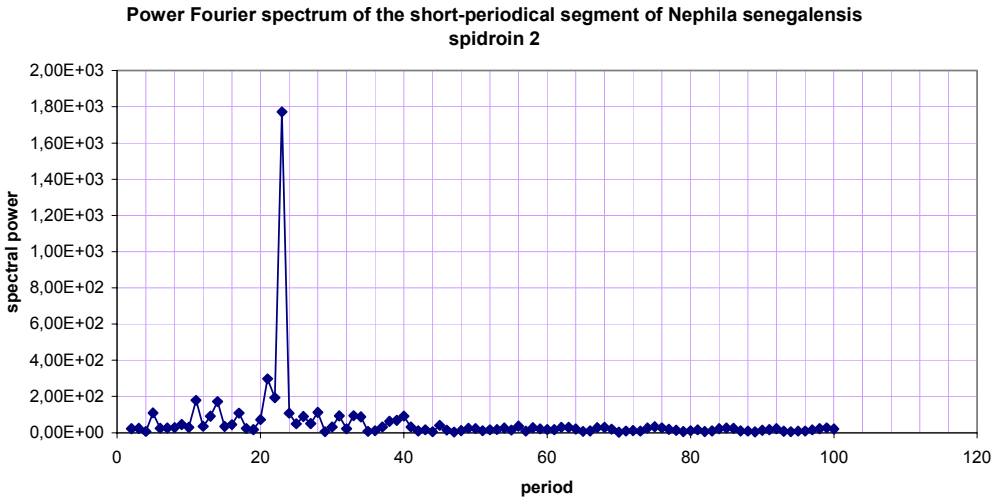


Fig. 2. Power Fourier spectrum of the short-period segment of *Nephelia madagascariensis* spidroin 2.

periodicities with periods 21, 23, and 25 amino acids. Then the whole sequence was scanned with the window of ~200 letters with step 24 and the spectral power of periods 21, 23, and 25 was monitored. There were identified three distinct domains with different dominant periodicities: (i) the 1382 residues long domain "with insertions"; (ii) the 571 residues long domain "without insertions"; and (iii) the 335 residues long "tail domain". We called the domains in such a way because some of the segments between poly-A within the domain "with insertions" contained an about 16 residue long insertion. The close inspection demonstrated that the domains "with insertions" could be approximated with two types of standard monomers, A and B with the whole segment having formula $\mathbf{ABA_2BA_5}$ where

A:

gpggygpgqqgpgaaaaasagr**gpggygpgqqgpggsgaaaaaa**grgpggygpgqqgpggpgaaaaaaagrgpggygpgqqgpgqqgpgg sgaaaaaaagrgpggygpgqqgpggpgaaaaagpggygpgqqgpgaaaaaaagsgpggygpgqqgpggpgaaaaaa

B:

gpggygpgqqgasaaaasaaagrgpggygpgqqgpggpsaaaagpggygpgqqgpsaaaaaagsgpggygpgqqgpggpgaaaaaagpggygp gqqgpgaaaaaaa

The domain "without insertions" can be approximated as C_6B, where

C: gpggygpgqqgpgaaaaasagrgpggygpgqqgpggsgaaaaaagrgpggygpgqqgpggpgaaaaaaa.

The tail domain has the formulae D_3 with

D:

gpgqqgpggygpgqq**gpggygpgqqgpsgpgsaaaaaaa**gpgqqgpggygpgqqgpggygpgqqgpsgpgsaaaaaaaagpgqqgpggyg pgqqgpggygpgqqgpsgpgsaaaaaaaaa.

Thus, the whole sequence can be approximated as $ABA_2BA_5C_6BD_3$, of which A,B,C, and D monomers have lengths of 166, 105, 69, and 123 respectively.

It should be noted that tail domain D demonstrates a strong similarity to domain without insertions A (shown with bold). Since the tail domain is known for many spider species, whereas there is only the one example of the sequence fragment long enough to have the domain "without insertions" read, it is tempting to speculate that we can reconstruct the sequence of the domain "without insertions" from the sequence of the tail domain. Although the basis for such speculation is less than solid (we've got only one example of a long fragment), the idea can be nevertheless promising, particularly in the case of spidroin I protein, for which there is no long enough fragments known at all.

Experimental verification. It is noteworthy that the basic structure of most of natural silks is α-helix or the helix of a collagen type. In this connection is important periodical patterns found in spidroin sequences are dramatically differ from those characteristic for α-helices and from collagen type helices. The periods found in spidroin agree with the β structure model. We tried to verify the β-structured model of spidroin arrangement with the help of FTIR and CD spectra. We studied specific features of secondary structures of recombinant proteins in solutions and in films. For the solution experiments it was found that CD-curve could be approximated by the left-handed helix of a polyproline II type and the β-structure. However, the share of the β-structure was not dominant. The films by recombinant proteins contained highly organized segments of antiparallel β-structure, segments of coils, and did not contain α-helix and polyproline II. IR-bands demonstrated the resemblance of secondary structures of the artificial protein, *Bombix mori* silk, and feather keratin.

Table. Components of the Amid I Band in the IR-spectra of some beta-structural proteins.

Component	Maximum frequency (cm-1)		
	Artificial silk	B.mori silk	Feather keratin
ν (II,0)	1631	1630	1634
ν (II,II)	1666		1665
ν (0,II)	1680	1691	1691
ν	1693		
ν2	1657	1656	1658

Acknowledgments

We wish to thank V.M. Lobachov and Ju.A. Lazavrev for there kind assistance in obtaining CD spectra.

References

1. Xu M., Lewis R.V. (1990) Structure of a protein superfiber: spider dragline silk. Proc. Natl Acad. Sci. USA. 87(18):7120-4.
2. Madsen Bo., Shao Z.Z., Vollrath F. (1999) Variability in the mechanical properties of spider silks on three levels: interspecific, intraspecific, and intraindividual. Int. J. of Biol. Macromol. 24:301–306.
3. Makeev V.J., Tumanyan V.G. (1996) Search of periodicities in primary structure of biopolymers: a general Fourier approach. CABIOS. 12:49–54.
4. Gatesy J., Hayashi C., Motriuk D., Justin W., Lewis R. (2001) Extreme diversity, conservation and convergence of spider silk fibroin sequences. Science. 291: 2603–2605.

CHARGE REPARAMETRIZATION FOR FAST ATOMIC-DETAIL CALCULATIONS IN PROTEINS

Schwarzl S.M., Huang D., Smith J.C., Fischer S.

IWR Biocomputing, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany, e-mail: sonja.schwarzl@iwr.uni-heidelberg.de

Key words: *proteins, atomic-detail calculations, Coulomb potential*

To simulate regulatory networks the rate constants of the reactions involved must be known. However, experimentally these are often difficult to determine. Atomic-detail computer simulations can in principle be used to estimate rate constants using force field methods combined with a quantum description of the reaction process. In such simulations, the electrostatics are represented by a Coulomb potential between partial atomic charges that are parametrized for small building blocks in vacuum and transferred to the macromolecule. In aqueous solution, however, these interactions are affected by the solvent polarization. While this can be described by numerically solving the Poisson-Boltzmann equation, it is computationally expensive. A procedure is presented to optimally reproduce the electrostatic potential in solution by reparametrizing the partial atomic charges in such a way that a simple Coulomb potential can still be used. The procedure allows to perform fast calculations of reaction processes in proteins while accounting for solvent screening effects.

CONFINEMENT MOLECULAR DYNAMICS AND ITS APPLICATION TO THE STUDY OF POTENTIAL ENERGY SURFACES AND CONFORMATIONAL TRANSITIONS IN BIOMOLECULES

Krivov S.V.*[1], [3]* *Chekmarev S.F.*, [1,2] *Karplus M.

[1] Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 67000 Strasbourg, France
[2] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, 02138 Massachusetts, U.S.A.
[3] Institute of Thermophysics, SB RAS, 630090 Novosibirsk, Russia
e-mail: chekmare@itp.nsc.ru
*Corresponding author

Key words: *biomolecule, potential energy surface, molecular dynamics*

Resume

Motivation: Systems of biological interest usually have potential energy surfaces (PESs) that consist of multiple minima separated by barriers that are large with respect to *kT* at normal temperatures. As a result, the behavior of the system becomes non-ergodic on the time scales accessible to standard molecular dynamics (MD) or Monte Carlo (MC) methods on current computers. It is thus useful to bias the search algorithm so that it violates the Boltzmann distribution but allows one to gain information about the system, including the knowledge of its PES, equilibrium properties and kinetics.

Results: A new, heuristic approach is presented, in which the MD trajectory of the system is successively confined to various basins of the PES. The approach is illustrated by a study of the solvated alanine-tetrapeptide, for which the various *cis* isomers are sampled starting with the all *trans* isomer, even though the barriers are over 25 kcal/mole. Comparisons with conventional MD are provided, confirming the efficacy of the approach.

Introduction

A knowledge of the PES of a system is fundamental to an understanding of its structural, thermodynamic and dynamic properties. Of particular interest at present are mesoscopic systems, such as proteins, nuclei acids and lipid membranes.

An approach for describing complex PESs based on topological mapping via disconnectivity graphs has been introduced recently (Becker, Karplus, 1997), the essential element of which is a knowledge of the local minima and the transition states connecting them. Their determination becomes increasingly expensive as the size of the system grows.

Most of the systems of interest have a PES that consists of regions of minima separated by barriers that are large with respect to *kT* at normal temperatures. The behavior of a system with such a PES is non-ergodic on the time scales accessible to standard MD or MC methods on current computers. It is, therefore, critical to develop methods that go beyond these methods. It clearly can be useful to bias the search algorithm so that it violates the equilibrium Boltzmann distribution. Many methods have been proposed for this purpose, which utilized various ways of deformation of the PES (as, the umbrella potentials; e.g., Bartels, Karplus, 1998) and elevating the temperature (as the multiple histogram methods; e.g., Labastie, Whetten, 1990). However, they are of limited value for efficient exploration of the PESs that superpose energy variations on very different scales, which is characteristic of biomolecules.

Approach

Here we present an alternative approach, which is based on a very simple concept of heuristic character that gives it a wide range of applicability. Specifically, various basins on the PES are sampled successively, with the choice of basins, after they have been visited at least once, determined by various possible strategies of surveying the PES (Krivov et al., 2002). To sample a specific basin, the approach makes use of the confinement technique (Chekmarev, Krivov, 1998; Chekmarev, 2001), which allows one to keep the MD trajectory of a system within the basin for an arbitrarily long time, and thus to sample the basin as thoroughly as desired. The approach offers essentially unlimited flexibility in the redistribution of the residence times of the system in different regions of the PES and provides a general framework for constructing optimal simulation schemes.

In its utilization with constant temperature MD simulations, such as the Langevin dynamics used here, the procedure is as follows. The system is placed in a certain basin and a MD run is begun at a temperature *T*. At regular intervals the system is quenched to check if the system (i.e. its representative point) is still in the given basin or has left it for another (connected) basin. If the system is in the original basin, the MD run is continued, but if it found to be in another basin, it is placed back into the original, and a new trajectory is initiated in this basin. Although the system does not leave the current basin for a time longer than the quenching interval, a record is kept of all the basins that are visited. It is thus becomes possible to

calculate not only the equilibrium properties corresponding to the given basin, but also the probabilities to pass into the connected basins, and thus to determine the kinetics of the system.

The choice of the next basins to be sampled depends on the goal of the study. For systems that are small enough so that a sampling of all (or a very large part) of the basins on the PES is possible, one can allow the system to pass into each new basin that has not been previously sampled and study it in turn. However, if the system is large, the goal may be to survey just a part of the PES (e.g., the low energy region that is important at ordinary temperatures), or to obtain "coarse" survey of the entire PES. In the first case, preference is given to the basins which are in the part of the PES of interest; in the second case, the basins investigated are chosen by the criterion that their minima differ most from those of the initial basin and other basins which have already to be sampled. Alternatively, one may be interested in the dynamics of the system, so that preference is given to basins with low energy transition states.

Results and Discussion

In the present study we mainly employed the first of the previously mentioned strategies that allows the system to pass into every new basin that has not been sampled previously, and applied it to an exploration of the PES of a tetrapeptide, the simplest peptides that can form a full α-helical turn. Specifically, we studied the solvated alanine-tetrapeptide.

The MD simulations were performed with the CHARMM program (Brooks et al., 1983), using the polar hydrogen parameter set for peptides and proteins (param19; Neria et al., 1996) and the ACS implicit solvent model (Schaefer et al., 1998). The friction coefficient in the Langevin equations was set equal to 64 ps^{-1}, and a time step of 1 fs was used. For quenching, a combination of steepest descent and the adopted-basis set Newton-Raphson minimization methods were employed (Brooks et al., 1983). To find the barrier between two minima, the TRAVEL algorithm (Fischer, Karplus, 1992) was used.

Figure presents some results of the work, related to the study of the PES landscape of the system. Fig.a shows the disconnectivity graph for a fragment of the PES that was sampled in confinement simulations for $5 \cdot 10^7$ timesteps (T=500K). During this time, 408 minima and 4800 transitions states were found. The graph reveals five clearly defined superbasins (funnels): the central one, which contains the α-helix and β-strand conformers, and four side funnels. The consideration of dihedral angles of the conformations associated with the minima has showed that the central funnel is associated with all *trans* conformation of the tetrapeptide, whereas the others represent mixed *trans*/*cis* conformations. Thus not only the manifold of states associated with the lowest energy all-*trans* peptide bond isomer are sampled, but also the *cis* peptide isomers, which are separated by high barriers from the all-*trans* form, are obtained without explicitly introducing them.

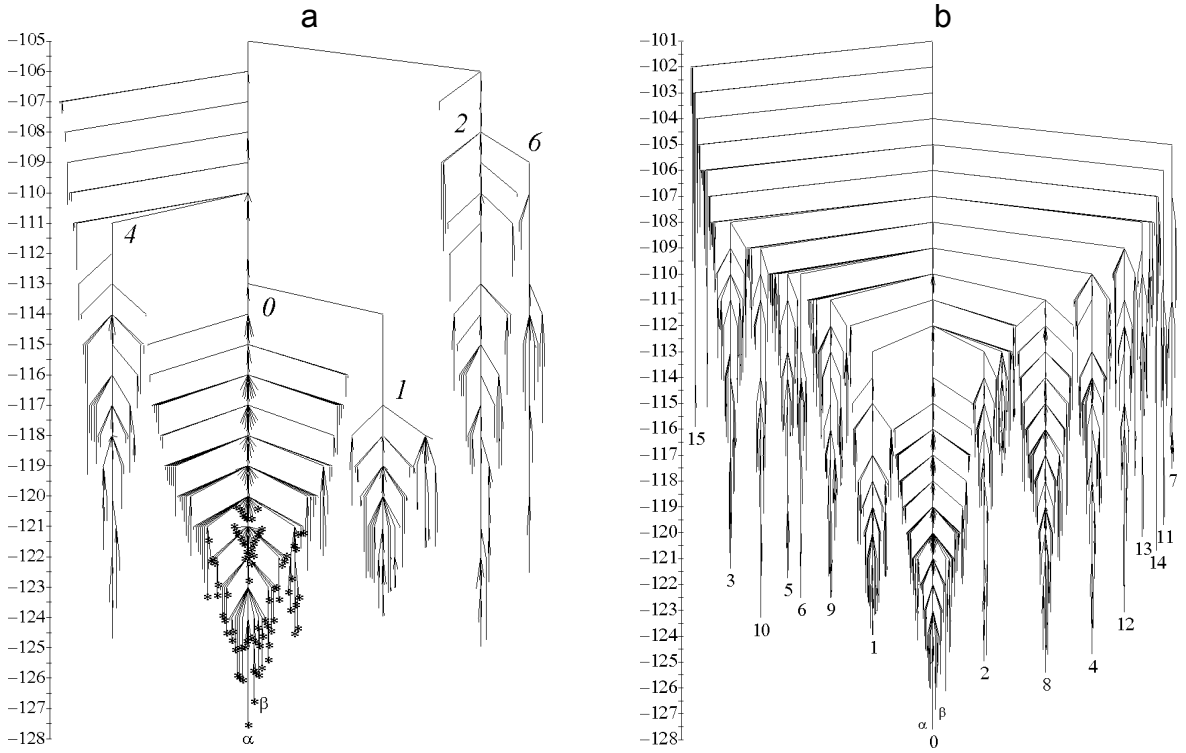


Fig. Disconnectivity graphs of the PES of the tetrapeptide. Energy level spacing is1kcal/mol. α and β indicate the minima corresponding to the α-helix and β-strand conformers, respectively. Conformers are numbered according to a reverse binary rule (i.e. the count is made from left to right), with 0 and 1 standing for the *trans* and *cis* isomer of the peptide group, respectively. For example, 7 = (1110) is the *cis-cis-cis-trans* conformation. ***a)*** Fragment of the PES; stars indicate the minima that were found in the conventional MD simulations. ***b)*** The entire coarse graining PES.

For the same lenght of the run, the conventional Langevin MD simulation was able to find only 97 minima and 670 transition states, all located in the lower part of the central funnel, where the MD trajectory was started (Fig.).

Confinement simulations done at a higher temperature (T=800K, $6.4 \cdot 10^6$ timesteps) have allowed us to perform a coarse graining survey of the entire PES (946 minima and 4930 transition states), Fig.b. The surface consists of 2^4=16 funnels that correspond to various *trans*/*cis* conformations of the peptide groups, from all *trans* (0) to all *cis* (15), with the minima of the funnels grouping into 5 bands of the minima depending on the number of *trans* and *cis* elements in the corresponding conformers.

Acknowledgment

The work was supported by the INTAS, grant № 2001-2126. The Laboratoire de Chimie Biophysique is partly supported by the CNRS, ISIS CR83, and by the Ministraire d'Education Nationale. S. F. Ch. acknowledges a support from the Siberian Branch of the RAS, grant № 65.

References

1. Bartels C., Karplus M. (1998) Probability Distributions for Complex Systems: Adaptive Umbrella Sampling of the Potential Energy. J. Phys. Chem. B 102, 865- 880.
2. Becker O.M., Karplus M. (1997) The Topology of Multidimensional Potential Energy Surfaces: Theory and Application to Peptide Structure and Kinetics. J. Chem. Phys. 106, 1495-1517.
3. Brooks B.R., Bruccoleri R.E., Olafson B.D., States D.J., Swaminathan S., Karplus M. (1983) CHARMM: A Program for Macromolecur Energy, Minimizatiom, and Dynamics Calculations. J. Comput. Chem. 4, 1187-217.
4. Chekmarev S.F., Krivov S.V. (1998) Confinement of the Molecular Dynamics Trajectory to a Specified Catchment Area on the Potential Surface. Chem. Phys. Lett. 287, 719-724.
5. Chekmarev S.F. (2001) Confinement Technique for Simulating Finite Many-Body Systems. In Atomic Clusters and Nanoparticles, Lectures at the Les Houches Summer School, Session No. LXXIII. Guet, G., Hobza, P., Spiegelman F., Gavid, F. (eds). Springer-Verlag and EDP Sciences, Les Ulis, 509-563.
6. Fischer S., Karplus M. (1992) Conjugate Peak Refinement: An Algorithm for Finding Reaction Paths and Accurate Transition States in Systems with Many Degrees of Freedom. Chem. Phys. Lett. 194, 252-261.
7. Krivov S.V., Chekmarev S.F., Karplus M. (2002) Potential Energy Surfaces and Conformational Transitions in Biomolecules: A Successive Confinement Approach Applied to a Solvated Tetrapeptide. Phys. Rev. Lett. 88, 038101.
8. Labastie P., Whetten R.L. (1990) Statistical Thermodynamics of the Cluster Solid-Liquid Transition. Phys. Rev. Lett. 65, 1567-1570.
9. Neria E., Fischer S., Karplus M. (1996) Simulation of Activation Free Energies in Molecular Systems. J. Chem. Phys. 105, 1902-1921.
10. Schaefer M., Bartels C., Karplus M. (1998) Solution Conformations and Thermodynamics of Structured Peptides: Molecular Dynamics Simulation with an Implicit Solvation Model. J. Mol. Biol. 284, 835-848.

PREDOMINANT CONFORMATIONS OF OLIGOPEPTIDE FRAGMENTS OF GLOBULAR PROTEINS

Vlasov P.K., Kilosanidze G.T., Ukrainskii D.L., Tumanyan V.G., Esipova N.G.

Engelhardt Institute of Molecular Biology, RAS, 119991, Moscow, Russian

***Key words**: secondary structure, conformation, left-handed helix of poly-L-proline-II type, olygopeptide frequencies, PDB*

Abstract

Motivation: The globular proteins constitute the major part of the structural protein databases. However, there is lack of data on the predominant conformations of the proteins' short fragments and the sequence specificity of the fragments. A special attention has to be drawn to the left helical conformation.

Results: Regions of the three main types (alpha-helices, beta-structures and left helices) of a secondary structure in globular proteins were studied throughout the PDB bank. The length and sequence of the fragments with regular conformation were analyzed. For the above types of secondary structure the sets of characteristic tetrapeptides have been revealed.

Contact: vlasov@imb.imb.ac.ru.

Introduction

The analysis of the occurrence of the three main types of a secondary structure have been undertaken on the basis of modern PDB bank release of 3-D structures of globular proteins (previous results were presented in (Vlasov et al.)). Dihedral angles φ and ψ for different individual amino-acid residues and short oligopeptides were determined during computations. Sample under investigation presents whole PDB bank or subset of non-homologous proteins. The subsets of structures corresponding to high-resolution were used also. Interestingly, the results for the subsets show close resemblance to those obtained for the total bank of structures.

Methods and Algorithms

The distribution of dihedral angle values for certain amino-acid residues was presented using the Ramachandran plot. The relative content for different types of conformations (namely the alpha-helix, the beta-structure and the left helix of the poly-L-proline-II type) was also estimated (see Fig. 1).

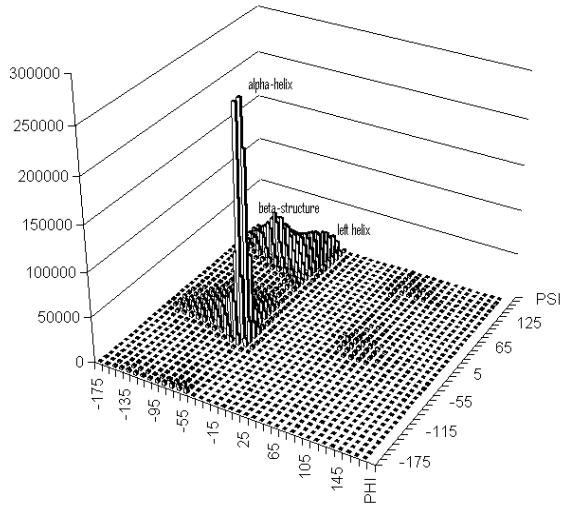


Fig. 1. Legend: Ramachandran plot for occurrence of single amino-acid residues in protein structures.

One of a reasonable approach to description of the stable conformations of fragments of a polypeptide chain includes using of idealized helix parameters (Adzhubei, Sternberg). Detailed analysis of such a kind was made in this work for PDB bank

structures. The considerable presence of a left-helix conformation in structures of globular proteins was established hand by hand to common regular structures.

Implementations and Results

For oligopeptides included two, three and four amino-acid residues, the frequency of their occurrence in each of the aforesaid secondary structure types was calculated. No trustworthy preference for one or other secondary structure type was found at the level of di- and tripeptides. Nevertheless tetrapeptides show some preference for distinct secondary structure (see α-helices tetrapeptides in Fig. 2).

We demonstrate at a first time the existence of tetrapeptides adopted left helical conformation of the poly-L-proline-II type in majority of cases which characterized by dihedral angles corresponding to the appropriate region on the Ramachandran map (see Fig. 3).

1	ANWM	98
2	RRCA	97
3	NWMC	97
4	QALW	97
5	INMV	97
6	MYLL	97
7	FISE	97
8	AMNK	96
9	RWYN	96
10	NMVF	96
11	CAKR	96
12	LANW	96
13	FNQD	96
14	SLRM	96
15	RMLQ	95
16	RMKD	95
17	DQLA	95
18	GHEQ	95
19	IERM	95
20	IKYL	95
	…	…

Fig. 2. Legend: column #1 – tetrapeptide number, column #2 - sequence, column #3 – occurrence percentage in α-helix

1	WHPK	71
2	SPQP	68
3	ETPS	58
4	PGPP	53
5	HPKA	59
6	PAQP	61
7	PQPG	60
8	PPGP	51
9	FVIR	59
10	APSP	55
11	PRPP	58
12	PPPP	45
13	PPQT	48
14	WKKD	57
15	RPEP	55
16	FPQR	50
17	PMAP	56
18	GPPD	49
19	YEPT	55
20	GPPG	46
	…	…

Fig. 3. Legend: column #1 – tetrapeptide number, column #2 - sequence, column #3 – occurrence percentage in left helix conformation of the poly-L-proline-II type.

Tetrapeptides "avoiding" the left-helix conformation were also identified. Those tetrapeptides in majority of cases characterized by dihedral angles beyond the region of the left helix of the poly-L-proline-II type (see Fig. 4).

1	AEKL	618
2	AKRV	602
3	WYNQ	503
4	VDAA	492
5	TGVA	488
6	ALEL	486
7	ELDK	485
8	ALLD	478
9	AKLK	459
10	DAAV	452
11	IGIG	445
12	ALIN	442
13	AIGR	440
14	KSEL	440
15	EKLF	436
16	AAHC	434
17	KELG	432
18	VDLL	431
19	GILR	426
20	EAEK	424
	…	…

Fig. 4. Legend: tetrapeptides definitely avoided left helix conformation of the poly-L-proline-II type; column #1 – tetrapeptide number, column #2 – sequence, column #3 – occurrence in PDB structures.

Statistical characteristics were calculated for the distribution of conformations of the residues within the tetrapeptide between the zones of the Ramachandran plot corresponded to the three principal types of secondary structures (the alpha-helix, beta-structure and the left helix of the poly-L-proline-II type).

Discussion

Since so short oligopeptides as four-member compounds exhibit distinct preference to the type of conformation, it is quite possible to draw useful statistics for analysis and subsequent prediction of the conformation type of peptide fragments.

Statistics of the sets of dihedral angles obtained in this work provides evidences that the length of a tetrapeptide is sufficient for the formation of characteristic and therefore predictable conformation. This property naturally is a consequence of concrete composition and sequence of the oligopeptide. The results obtained imply that the electrostatic properties of such a long peptide affect the conformations.

Acknowledgments

This work was supported in part by the Russian Foundation for Basic Research (Grants 00-04-48351 and 02-04-49114) and INTAS 99-1476.

References

1. Vlasov P., Kilosanidze G., Ukrainskii D., Kuzmin A., Tumanyan V., Esipova N. (2001) Left-handed Conformation of Poly-L-proline-II Type in Globular Proteins. Sequence Specificity, Biophysics. 46-3, 573-576.
2. Adzubei A., Sternberg M. (1993) Left-handed Polyproline II Helices Commonly Occur in Globular Proteins. J. of Mol. Biol. 229, 472-493.

RESOURCES FOR THE ANALYSIS OF PROTEIN SEQUENCES AND STRUCTURES IN THE GENEEXPRESS SYSTEM

Afonnikov D.A., Ivanisenko V.A., Grigorovich D.A., Valuev V.P., Nikolaev S.V., Kolchanov N.A

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

Key words: *genome, protein, structure, amino acid sequences, database, integration*

Motivation: Proteins play an important role in the processes of cell active life. Possessing a unique structure, they perform many different functions: participate in the regulation of gene expression, in the biosynthesis and degradation of the molecular components of living organisms, catalyze biochemical reactions occurring in living systems. Information on the peculiarities of protein structure, their interactions with DNA, RNA and other compounds is necessary for solving a wide range of problems of pharmacology, genic and protein engineering, biotechnology.

Results: A series of databases and programs combined into a general module Protein Integration Level was developed to solve the problems of the analysis of protein structure, function and evolution in the framework of the integrated system GeneExpress 2.1. This module includes the databases of the extended annotation of PDB structure, active protein sites and sequences resulting from artificial selection. It also contains the programs for the analysis of coordinated substitutions of residues in proteins.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/protein,shtml

Introduction

A study of genome function patterns is one of the basic problems of the present-day molecular biology. A particular feature of a genome is the high intricacy of its organization which can be described as a hierarchy. A complete and adequate investigation of genome function is impossible without studying carefully all levels of its hierarchy. One of these levels is a protein level related to the synthesis, function, and degradation of protein macromolecules.

One of the systems realizing the integrated approach to the study of genome regulation is the GeneExpress system (Kolchanov et al., 1999). To study genome function in this system at a protein level, we have developed the integrated module Protein Integration Level.

This paper outlines the databases and programs included in the integrated module Protein Integration Level of the system GeneExpress 2.1.

RESOURSES CONTAINED IN THE MODULE PROTEIN INTEGRATION LEVEL IN THE COMPUTER SYSTEM GENEEXPRESS 2.1.

Resources involved in the module Protein Integration Level of the computer system GeneExpress 2.1, include the databases of protein structures EnPDB, PDBSite, sequences (ASPD), the programs for the analysis of structures (PDBSiteScen) and amino acid sequences (GRASP) (see Table). This module can be found in http:wwwmgs.bionet.nsc.ru/mgs/gnw/protein.shtml. The main peculiarities of these databases and programs are shown below.

Table. The list of resources involved in the module Protein Integration Level of the computer system GeneExpress 2.1.

Resource	Brief description	Address in Internet
Database EnPDB	System of extended indexation of PDB records	http://wwwmgs.bionet.nsc.ru/mgs/gnm/enpdb/
Database PDBSite	Information on the structure and pysico-chemical properties of active protein sites	http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/
Program PDBSite Scan	Search for active sites in protein structures by the sequence and spatial distribution of residues	http://wwwmgs.bionet.nsc.ru//mgs/sysems/ fastprot/pdbsitescan.html
Database ASPD	Information on peptide and protein sequences obtained by in vitro selection	http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd
Program CRASP	Analysis of coordinated substitutions in protein families	http:/wwwmgs.bionet.nsc.ru/mgs/gnw/crasp

Database EnPDB. A complete analysis of relationship between the spatial structure of protein and its sequence is impossible without using genetic, evolution, and biochemical data. Therefore, when analyzing proteins, an important problem on the integration of structural data with other sources of biological information arises (Weissig, Bourne, 1999). On the other hand, the problems on the comparative analysis of protein structures call for the development of powerful and flexible methods for creating the samples of the 3D structures of proteins with certain characteristics (the sites of the same type, similar structural characteristics, etc.). These problems cannot be solved without developing the system of the extraction and indexation of information kept in the PDB database (Berman et al., 2000). The goal of the creation of the EnPDB system was to extend the potentialities of the indexed search for information in the records of the PDB database (Grigorovich, Ivanisenko, 2002). This extension was first concerned with the information contained in the heading part of the record. In this case, a series of fields containing nonuniform information were splitted into several blocks. Thus, for example, the field HEADER was splitted into three areas: ID, HEADER and DATA. The areas containing information on the peculiarities of protein structure and molecular complexes were added. These included the number of α-helices, β-strands, polypeptide chains, nucleic acids, heteroatoms, etc. The number of references to the databases containing additional information on the structure and function of biological micromolecules was substantially increased. The EnPDB interface was realized using the SRS system (Zdobnov et al., 2002). The extended indexation enabled both the search with the help of criteria for structural characteristics and the formation of the samples of proteins, nucleic acids or their complexes satisfying these criteria. For example, requesting the fields *HelixAmount*, *SheetAmount*, the proteins can be chosen whose structures contain α-helices and no β-sheets.

Database PDBSite. Protein function as an entity of genetic network depends on both its interaction with other entities of this network (local protein function) and its role in the genetic network as a whole (integral protein function) (Karp, 2001). Since interaction between the protein and the other components of genetic network is determined by a set of its active sites, the local protein function cannot be studied without studying the function of its active sites. Information on the active protein sites is the focus of attention of our database PDBSite (Ivanisenko et al., 2002a). This database was formed using the annotation of protein sites available in the field SITE of the PDB records. Each record of the PDBSite database corresponds to a concrete function site of a concrete protein structure. For active sites we also calculated their structural and physico-chemical characteristics (the means, the sum and the space moment of a series of physico-chemical properties, surface area, available for a solvent, the coordinates of a mass center), the coefficient of discontunuity by the primary structure, etc. At present, 4723 sites are annotated in the PDBSite database. The database interface is realized through the SRS (Zdobnov et al., 2002).

The PDBSiteScan program for search of the space structure of protein function sites. This program is used to search for active sites in protein space structures by comparing the types of amino acid residues forming an active site with their spatial distribution (Ivanisenko et al., 2002b). The program provides the list of sites from the PDBSite database found in the protein tested.

Database ASPD. The up-to-date achievements in molecular biology and biotechnology allowed the creation of a number of novel techniques for studying biological macromolecules that are widely used for solving the problems of DNA, RNA, and protein analysis. One of these technologies is the method of selection of peptides and proteins *in vitro* (Roberts and Ja, 1999). As a result of this experiment, up to several tens of sequences specifically binding to certain substrates are chosen from the pool of sequences containing all possible variants of amino acids in their positions. As a rule, their set is represented as a multiple alignment. To record such information, we have developed the ASPD database (Artificially Selected Proteins/Peptides Database) (Valuev et al., 2001). Each entry of this database contains a sequence alignment, experimental details, reference to both the total sequence of the protein analyzed and the paper from which this alignment was taken. This database provides additional information obtained by analyzing the sets of alignment. This is the matrix of amino acid substitution and information on the binary correlations of the values of physico-chemical characteristics of amino acid residues in alignment positions. The interface of the ASPD database allows one to search for sequences using the BLAST program (Altschul et al., 1997). Thus, the information provided by the ASPD database can be used to solve a wide range of problems on the analysis of the peculiarities of the function of active protein sites.

The program pack CRASP for the analysis of the coordinated substitutions of aminoacids

The study of the sets of the homologous sequences of isofunctional proteins is one of the most important methods of analysis in molecular biology. It is assumed that the protein function and structure remain almost constant upon evolution. Thus, the physico-chemical protein characteristics determining a specific packing of a polypeptide chain and the function peculiarities upon evolution should also be maintained at a constant level. The detection of such conserved characteristics upon analysis of homologous sequences can substantially add to our knowledge about the function, structure and evolution of the proteins under study. One of the possible ways of keeping these characteristics constant are the coordinated substitutions of amino acid residues. To study the peculiarities of the coordinated evolution of amino acid residues, we have developed the CRASP program package (Afonnikov et al., 2001; Afonnikov, 2002). This program package can be used to reveal and analyze not only the substitutions of amino acid residues in proteins occurring in a correlated manner but also the different integral physico-chemical characteristics of protein sequences. It is based on both the method for detecting and analyzing correlations among the values of the physico-chemical characteristics of residues in the positions of protein sequences and the method for analyzing the conserved integral physico-chemical protein characteristics. The package

consists of two groups of progams for: (1) the analysis of binary correlations among amino acid substitutions, and (2) the analysis of the integral characteristics of the groups of protein positions.

Thus, the module Potein Integration Level of the computer system GeneExpress 2.1 makes it possible to solve a series of important problems on the analysis of protein sequences and structures.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (№ 01-07-90376, 01-07-90084), Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Projects № 65), National Institute of Health USA (№ 2 RO10HG-01539-04A2), the Department of Energy USA (№ 535228 CFDA 81.049).

References

1. Afonnikov D.A. (2002) Analysis of the contribution of co-adaptive substitutions to the constancy of the physico-chemical characteristics of binding sites of akp proteinkinase. (This volume).
2. Afonnikov D.A., Oshchepkov D.Y., Kolchanov N.A. (2001) Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with coordinated substitutions. Bioinformatics. 17, 1035-46.
3. Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Miller W., Lipman D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25**,** 3389-3402.
4. Grigorovich D.A., Ivanisenko V.A. (2002) EnPDB – PDB database search system. (This volume).
5. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Wessig H., Shindyalov I.N., Bourne P.E. (2000) The protein data bank. Nucl. Acids Res. 28**,** 235-242.
6. Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A. (2002a) PDBSite – database on protein active sites and their environment. (This volume).
7. Ivanisenko V.A., Debelov V.A., Pintus S.S., Matsokin A.M., Nikolaev S.V., Grigorovich D.A., Kolchanov N.A. (2002b) PDBSiteScan" a tool for the search of bestmatching superpositions by PDBSite databse. (This volume).
8. Karp P.D. (2000) An ontology for biological function based on molecular interactions. Bioinformatics. 16, 269-85.
9. Kolchanov N.A., Ponomarenko M.P., Frolov A.S., Ananko E.A., Kolpakov F.A., Ignatieva E.V., Podkolodnaya O.A., Goryachkovskaya T.N., Stepanenko I.L., Merkulova T.I., Babenko V.V., Ponomarenko Y.V., Kochetov A.V., Podkolodny N.L., Vorobiev D.V., Lavryushev S.V., Grigorovich D.A., Kondrakhin Y.V., Milanesi L., Wingender E., Solovyev V., Overton G.C. (1999) Integrated databases and computer systems for studying eucaryotic gene expression. Bioinformatics. 15, 669-86.
10. Roberts R.W., Ja W.W. (1999) *In vitro* selection of nucleic acids and proteins: What are we learning? Curr. Opin. Struct. Biol. 9, 521-9.
11. Valuev V.P., Kuropatov D.A. (2000) Automatic generation of recognition programs for amino acid sequences. Computationl technologies. 5, special issue, 67-74.
12. Valuev V.P., Afonnikov D.A., Ponomarenko M.P., Milanesi L., Kolchanov N.A. (2002) ASPD (Artificially Selected Proteins/Peptides Database): a database of proteins and peptides evolved in vitro. Nucl. Acids Res. 30, 200-202.
13. Weissig H., Bourne P.E. (1999) An analysis of the protein data bank in search of temporal and global trends. Bioinformatics. 15, 807-831.
14. Zdobnov E.M., Lopez R., Apweiler R., Etzold T. (2002) The EBI SRS server-recent developments. Bioinformatics. 18, 368-73.

ENPDB: A RETRIEVAL SYSTEM FOR THE PDB DATABANK

Grigorovich D.A., * Ivanisenko V.A

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

Key words: *PDB, tertiary structures, SRS, databases*

Summary

Motivation: Data on the spatial structures of DNA, RNA, and proteins accumulated in the Protein Data Bank (PDB) are of paramount importance for medicine, biology, and biotechnology. The huge volume of PDB annually increases following a geometric progression. At present, it has over 6 Gbytes. Efficient search over this volume demands use of modern computer methods.

However, the PDB structure was developed in the infancy of computer technologies for searching databases. This database contains much poorly formalized information. This reduces significantly the efficiency of search and analysis of the spatial structures of biological macromolecules.

Results: A converter program has been developed for formalization of PDB, which would allow broad search over the database field. The EnPDB database has been constructed with the use of the system SRS. The database is integrated with a system for visualization of spatial structures and other databases.

Availability: http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+ENPDB

Introduction

Data on the spatial structures of DNA, RNA, and proteins are accumulated in the Protein Data Bank (PDB) (Bernstein et al., 1977). This bank is the only official source of research information on the known spatial structures of macromolecules. It contains information on the amino acid sequence of a protein, also called primary structure (sequence of characters), the secondary structure (local spatial folding of polypeptide chains into alpha helices and beta turns), active sites of proteins, coordinates of atoms forming the protein, etc. A typical protein contains thousands of atoms, and the X, Y, Z coordinates are stored for each of them. For nonprofitable establishments performing basic studies, this information is available by Internet. However, when PDB was created, there were few decoded structures. Therefore, the problem of automated retrieval and read-out of information on spatial structures was not urgent. The implementators of PDB put emphasis on the complete representation of structural information, experimental methods and conditions, and functional features of the macromolecules. The structure of the base itself was improved with regard to primary accumulation of data and the input of new information. For example, information on various macromolecules in PDB is stored in individual files. There are also reserved fields, which were assumed to allow "evolution" of the base following the evolution of experimental and theoretical methods for obtaining new knowledge.

An abrupt change in methods for synthesis, isolation, and crystallization of biologic macromolecules has occurred in recent years. As a result, the number of decoded structures has exceeded 16,000 and continues to grow, doubling each year. Efficient search throughout this huge volume of information can be performed only with the use of modern computer technologies. One of such technologies is SRS (http://srs.ebi.ac.uk/). Its efficient use depends directly on the degree of data formalization. Therefore, one of the main tasks was transformation of the PDB bank to a more formalized state. This task included the following objectives: (1) classification of information stored in PDB; (2) selection of significant information; (3) construction of a conversion algorithm; and (4) implementation of the algorithm in the form of a converter program.

An additional task was the development of links of the EnPDB base with other databases and addition of new information, which would enlarge the possibilities of the search.

One of the difficulties in using information on the spatial structure of a macromolecule is that it cannot be understood without auxiliary software. The most commonly known package of the freeware means for visualization of spatial structures is RasMol. With its help, a user can visualize a single protein on his computer after its read-out from PDB. Thus, the next task was to integrate EnPDB with the visualization program RasMol.

Methods and Algorithms

The structure of EnPDB was developed so that most data could be indexed with SRS tools. During conversion of PDB to EnPDB, some PDB fields were decomposed into a series of new fields bearing uniform information. For example, the field HEADER was divided into three fields: ID, HEADER, and DATA.

New information was also added. We introduced new fields, characterizing the structural features of macromolecules calculated from data on tertiary structure: the number of alpha helices, beta turns, polypeptide chains, nucleic acids,

heteroatoms, etc. More references to databases containing additional information on the structure and function of biological macromolecules were added. The fields are listed in Table 1.

Table 1. Fields of the bank EnPDB in the system SRS.

Name	Short Name	Type	No. of Keys	No. of Entry
ID	id	id	16777	16777
Header	hdr	index	1117	31998
Date	dte	num	3355	16777
Title	ttl	index	11455	102564
Compound	cpd	show	0	0
Molecule	mol	index	6736	38259
Synonym	snm	index	3730	15097
EC	ec	index	834	6225
BioUnit	bun	index	347	4068
Gene	gen	show	0	0
MolSource	mls	index	4	12360
Source	src	index	6130	58526
Synthesis	snt	index	3165	42545
Keyword	kw	index	6320	75537
Technique	tch	index	39	25668
Author	aut	index	11666	69832
Jrnl	jrn	show	0	0
JrnlAuthor	jau	index	15644	94468
JrnlTitle	jti	index	12750	227024
JrnlRef	jre	index	608	21556
JrnlVolume	jvo	index	521	20923
JrnlYear	jye	num	39	19333
Remark_1	jrn	show	0	0
Resolution	res	real	257	10110
ChainAmount	cha	num	30	12350
ChainSizes	chs	num	763	16564
HelixAmount	hla	num	197	16777
SheetAmount	sha	num	181	16777
DnaRnaAmount	dra	num	11	12350
ProteinAmount	pra	num	29	12350
HetAmount	hta	num	99	12360
Heterogen	htg	index	5359	48112
LinkEmbl	lem	index	7196	24870
LinkPir	lpi	index	3047	9529
LinkSwissProt	lsw	index	3188	8320
LinkTransfac	ltf	index	100	190
LinkTrrd4	lt4	index	173	794

The algorithm for conversion of PDB was implemented in the Perl language. The original entries of the PDB database belong to two classes: the old format and formats 2.0, 2.1, 2.2, and 2.3. Old entries are poorly formalized and are processed with a single subroutine. New entries are better formalized and are processed with a subroutine that uses more fields from a PDB entry. At present, the number of new-format entries is three times as great as that of old formats. The conversion generates the file enpdb.dat, which is actually the base EnPDB.

The resulting file enpdb.dat is used in the system SRS. For this purpose, the entry parser was written in the language Icarus.

Implementation and Results

The releases of EnPDB of 2000 and 2002 are compared in Table 2. The comparison shows that the proportions of indices did not change except for the number of records with heterogens, whose content increased from 39 to 52%. This implies that scientists have crystallized more proteins with heterogens in the last two years than before.

Table 2. Comparison of two EnPDB releases.

	Release 2000	Release 2002
Entries	9902	16777
Monomeric proteins	2633	5704
Polymeric proteins	2289	5337
Protein-nucleic acid complexes	696	1309
Entries containing only nucleic acids	375	714
Entries with heterogens	3858	8762

The data bank EnPDB (http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+ENPDB) is integrated with the commonly known data banks TRRD, Swiss-Prot, Pir, Transfac, EMBL, GeneBank, etc. These bases are available from EnPDB by means of hypertext references with the use of SRS facilities. We also implemented access to PDB allowing

users to retrieve information not included in EnPDB yet. Besides, the RasMol viewer can be called for visualization of spatial structures. An example of visualization is shown in Fig.

Indexing of structural information with SRS facilities allowed search with using criteria of structural features and construct samples of proteins, nucleic acids, or their complexes meeting these criteria. For example, operation with the field combination ProteinAmount, DnaRnaAmount (amount of nucleic acids), and HetAmount (number of heteroatoms) allows construction of queries for search of all records containing spatial structures, monomeric protein molecules, proteins of several subunits, protein–protein complexes, protein–nucleic acid complexes, or protein–ligand complexes. Construction of samples of this sort is one of the most laborious tasks in the analysis of protein structure and function.

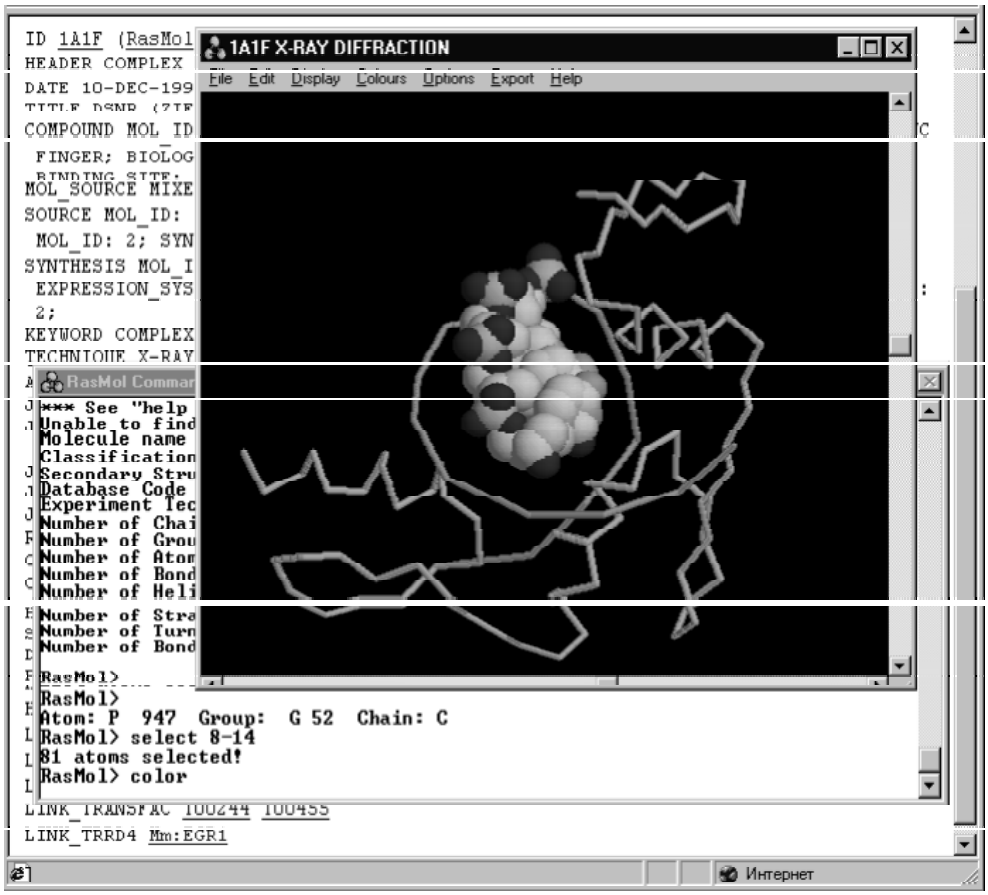


Fig. Visualization of an EnPDB entry with the help of the RasMol viewer.

For example, we want to find DNA/RNA protein complexes containing ligands MAGNESIUM, ZINC, or INOSIN and, in addition, meet the following requirements: the number of protein molecules no less than 2, the number of chains in protein molecules no less than 2, chain length in proteins no less than 50, the number of alpha helices no less than 2, and the number of beta turns no less than 3. We address the extended search page in the base EnPDB and properly complete the fields. For example, we select the sign ">=" in the field ChainSizes and enter 50, and so on. After clicking "Submit Query", we obtain a list of entries in the EnPDB database meeting the specified requirements. The resulting sample can be further investigated to reveal any regularities.

Discussion

With the volumes of information currently stored in PDB, the development of subordinate databases and their integration with PDB is an important task (Weissig et al., 1999, Berman et al., 2000). A subordinate database would considerably enhance the ability of the bank to search for spatial structures of protein with specified properties. Use of EnPDB contributed much to the development of software packages and databases, such as PDBSite (Ivanisenko et al., 2002), PDBSiteScan (Ivanisenko et al., 2002), and CRASP (Afonnikov, 2000).

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); US Department of Energy USA (grant № 535228 CFDA 81.049).

References

1. Afonnikov D.A. (2000) CRASP: a software package for analysis of physicochemical parameters of aligned sequences of protein families. II International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000). Novosibirsk, August 7-11. 2, 145-148.
2. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. Nucl. Acids Res. 28, 235-242.
3. 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. (1977) The Protein Data Bank: a computer based archival file for macromolecular structures. J. Mol Biol. 112, 535-542.
4. http://srs.ebi.ac.uk/
5. http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+ENPDB
6. Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A. (2002) PDBSite: database on protein active sites and their environment. This volume.
7. Ivanisenko V.A., Debelov V.A., Matsokin A.M., Pintus S.S., Grigorovich D.A., Kolchanov N.A. (2002) PDBSiteScan: a tool for search for the best-matching superposition over PDBSite. This volume.
8. Weissig H., Bourne P.E. (1999) An analysis of the Protein Data Bank in search of temporal and global trends. Bioinformatics. 15, 807-831.

PDBSITE: A DATABASE ON PROTEIN ACTIVE SITES AND THEIR ENVIRONMENT

[1,2] Ivanisenko V.A., [1] Grigorovich D.A., [1] Kolchanov N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk region, Russia
e-mail: salix@bionet.nsc.ru
*Corresponding author

Key words: *biologically active protein sites, protein tertiary structure, databases*

Resume

Motivation: The database Protein Data Bank (PDB) contains data on biologically active sites of many proteins: ligand-binding domains, enzyme catalytic centers, sites experiencing biochemical modification, etc. However, these data are of very limited access in the modern retrieval systems. Development of a database containing information on features of active sites and their spatial environment would provide a basis for comprehensive study of the properties of such sites.

Results: We have constructed the database PDBSite on biologically active sites retrieved from the database PDB. PDBSite contains description of site functions; lists of residues and their positions; structural features, calculated from 3D structures of the proteins; and physicochemical features of the sites and their spatial environment. The relationships between the properties of the residues of the sites and the residues of their environment have been analyzed.

Availability: http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+LibInfo+-newId+-lib+PDBSite or http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/

Introduction

Information on biologically active sites is of paramount importance in solving many problems of molecular biology, biotechnology, and medicine. Highly specific biological activities of proteins are provided by the unique structure of active sites and their environment. For example, catalytic centers of enzymes occur in cavities (Chothia, 1976), and antigenic determinants are shaped as protrusions on a protein surface (Davies, Cohen, 1996). The structure of sites is largely dependent on their environment. The activity of many natural and mutant proteins depends on the physicochemical properties of the residues surrounding the functional sites (Ivanisenko, Eroshkin, 1997). Data on the 3D structures of proteins are necessary for determination of the spatial environment of their biologically active sites. The database PDB (Bernstein et al., 1997) contains the data on 3D structures of proteins. For many proteins, the amino acid residues composing biologically active sites are marked, and the sites are briefly described. The purpose of the present study is the development of a daughter database PDBSite on the specific features of spatial organization of the biologically active sites stored in PDB and their environment.

For this purpose, we developed original methods, algorithms and software for calculating the 3D environment of sites and their structural and physicochemical properties.

Most investigations of the structure–function organization of functional sites are aimed at the invariant properties of these sites or their environment (Bagley, Altman, 1995; Sekharuda, Sundaralingam, 1988). We performed search for correlations among the variable properties of sites and their environment. This showed that a correlated change of the property pairs "site–environment" is a typical feature of the sites. These results may contribute to understanding the mechanisms of site operation and evolution.

Materials and Methods

The PDBSite database includes data obtained by processing of the following PDB fields: HEADER, TITLE, KEYWDS, REMARK 800, SITE, and ATOM. Grammar analysis programs were used for processing of PDB records. If a single PDB record contained data on several sites, individual records were created for each site in PDBSite. The Internet access to PDBSite was performed with the use of the Sequence Retrieval System (SRS).

The spatial environment of sites was calculated as follows. The least parallelogram including all the atoms of the amino acid residues of a site was constructed from the coordinates of the protein atoms. The spatial environment of the site was assumed to include each amino acid residue such that at least one of its atoms was within the parallelogram.

The following properties of sites and environment were calculated: solvent accessibility of amino acid residues; the mean value, sum, and spatial moment of amino acid physicochemical indices; coordinates of the geometrical center of mass for each residue; and the pairwise distances between the centers of mass of the residues. Another site index calculated was its discontinuity in the primary structure. The spatial moment was calculated as follows:

$$SM = \left\{ \left[\sum_{i=1}^{N} p_i x_i \right]^2 + \left[\sum_{i=1}^{N} p_i y_i \right]^2 + \left[\sum_{i=1}^{N} p_i z_i \right]^2 \right\}^{1/2},$$

where p_i, (i = 1, 2, ...,N) is the value of a certain property of the ith residue of the 3D site, comprising N amino acids; x_i, y_i, z_i are coordinates of the Ca atom of the ith residue taken with reference to the geometrical center of the 3D site.

The discontinuity index of a site was taken to be $\frac{1}{N}\sum_{i=1}^{N}\left(P_{i+1} - P_i - 1\right)$, where N is the number of residues of the site and P_i is the ordinal number of the ith residue of the site in the protein sequence. This index reflects the mean number of positions between the neighboring residues of the site in the primary structure. For example, if a site consists of a continuous sequence fragment, then the discontinuity index is zero. To calculate the solvent accessibility of amino acids, we used the program DSSP (Kabsch, Sander, 1983).

Results and Discussion

The fields of the PDBSite database used for search in the system SRS are listed and briefly described in Table 1.

Table 1. The description of PDBSite fields that can be queried.

Field name	Description
ID	Entry identifier.
PDBID	PDB ID code.
Header	PDB classification for the entry. The field content corresponds to that in PDB.
Title	Title for experiment or analysis described in the entry. The field content corresponds to that in PDB.
Keyword	Keywords describing the macromolecule. The content corresponds to that of KEYWDS in PDB.
Molecule	Contains names of macromolecules from the COMPND of PDB and is designed to look for entries by the names of macromolecules.
NumSiteChains	Number of different chains to which the residues of the site belong.
SiteDescr	Description of the site. The content corresponds to that of SITE_DESCRIPTION subfield of REMARK 800 field of PDB.
ResidueNotAA	Names of residues that are not amino acids but are present in the site.
LenSite	Number of residues in the site.
LenSurround	Number of residues in the site environment.
ExposureSite	Average exposure of residues in the site.
ExposureSurround	Average exposure of residues in the site environment.
Discontinuity	Discontinuity of the site according to its primary structure.

In addition, the database contains data on the structural and physicochemical indices of sites and their spatial environment that are not used for search but can be applied to analysis of the structure-functional organization of sites.

The PDBSite base contains the descriptions of 4723 sites. For statistical analysis, we obtained a nonredundant set by exclusion of complete analogs related to protein duplication in PDB. This set included 4038 sites. All the sites can be divided into three groups: continuous, discontinuous, and such discontinuous sites whose residues correspond to different subunits of a molecule. The proportion of the last group was considerable - 10.6%. The distribution of sites for the discontinuity index is shown in Fig. 1.

The figure shows that most sites present in the database are discontinuous. Thus, the patterns based only on the primary structure can hardly be developed for many sites. Recognition of the residues of such sites requires at least consideration of the matrix of distances between protein residues in the tertiary structure. We found that sites with increased discontinuity have limited solvent accessibility (Fig. 2); i.e., burred residues form highly discontinuous sites.

For further analysis, sites were grouped according to their functions. Only those sites whose function is unambiguously described in the field SiteDescr were taken into consideration. Of them, we chose site types including no less than 10 members. Thus, 3611 sites of 30 types were examined. Figure 3 shows the hierarchical classification of sites of different types for their mean amino acid composition. The sites clustered into two major groups in the resulting tree. One of them is dominated by organic ligand-binding sites, and the other, by metal ion-binding sites. Sites of acetylation fell into the first group, and the sites of glycosylation and phosphorylation, to the second one. Catalytic centers of enzymes belong to the first group. Both groups contain also local clusters of sites of different types. In general, the tree constructed according to the amino acid composition of sites is in good agreement with their function.

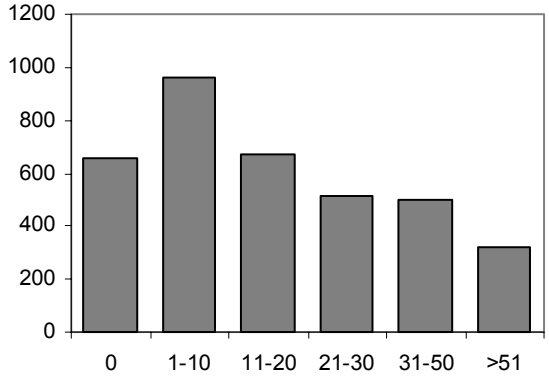


Fig. 1. Distribution of sites for discontinuity.

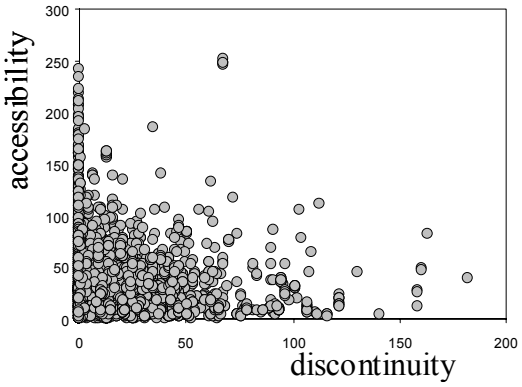


Fig. 2. Interrelation between the discontinuity and solvent accessibility of sites.

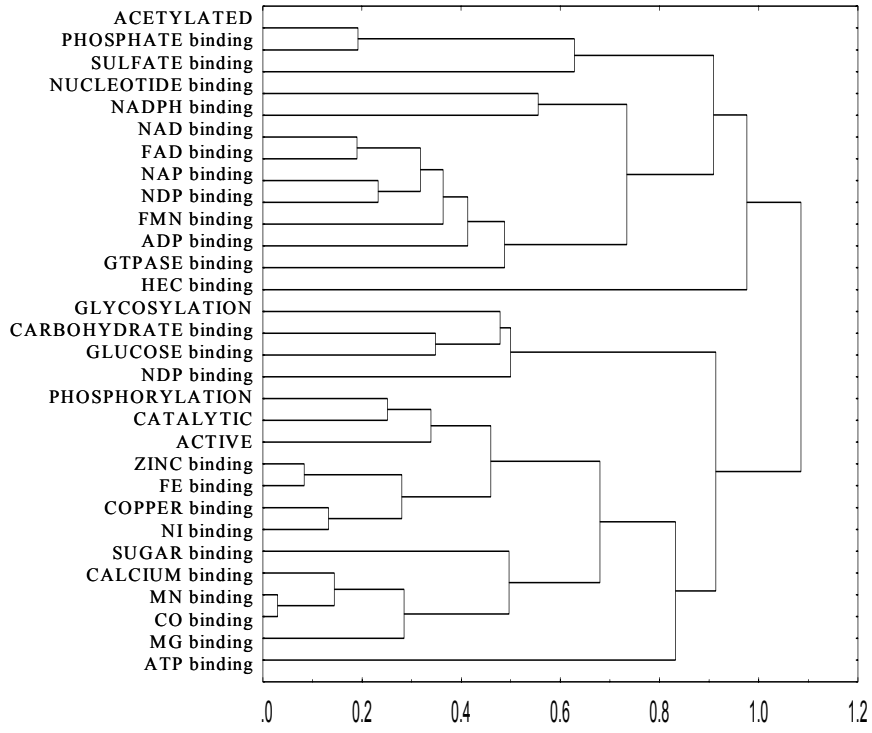


Fig. 3. The hierarchical tree classifying the sites of various types according to their amino acid composition.

Features presented in the PDBSite database were used for analysis of correlations between the physicochemical properties of sites and their environment.

Figure 4 illustrates the correlation between the hydrophilicity of the Mn-binding, Co-binding, and Mg-binding sites, pooled into one group, and the hydrophobicity of their solvent-exposed environment, i.e., residues on the molecule surface. It is apparent that the hydrophobicity of the exposed environment increases with site hydrophilicity.

The results of our analysis suggest that the function of site environment manifests itself as early as the stage of initial recognition of the site target and correct orientation of the site with respect to the target.

The mean or total physicochemical properties of sites were found to correlate with their spatial moment. This can be related to the importance of the distribution of these properties over the spatial structure of the site.

In particular, the increase in the number of positively charged residues in ATP-binding sites correlates with the increase in charge moment (Fig. 5). To put it differently, residues are aggregated into clusters of positively and negatively charged ones.

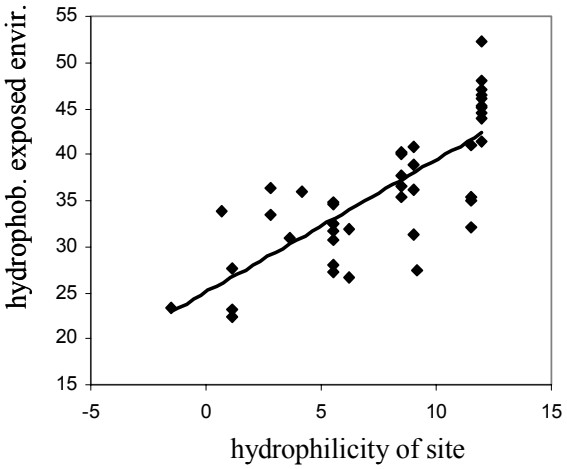


Fig. 4. Correlation ($R = 0.77$) between the hydrophilicity of the Mn-binding, Co-binding, and Mg-binding sites, bulked into one group, and the hydrophobicity of their exposed environment; $P>0.95$.

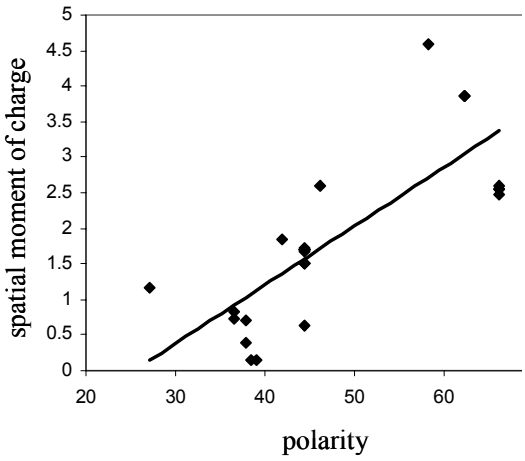


Fig. 5. Correlation ($R = 0.78$) of the polarity of ATP-binding sites and the spatial charge moment. The correlation is significant at $P>0.95$.

Conclusions

We have developed the PDBSite database, which can be applied to various tasks in the investigation of the functions of proteins and their active centers.

The PDBSite database is equipped with an automated system for search for structural similarity between active sites and a user-defined 3D structure of a certain protein (Ivanisenko et al., 2002). The search for structural similarity over PDBSite allows recognition of active sites in the spatial structure of proteins.

Our results indicate that the correlations between physicochemical properties of sites and their spatial environment are characteristic of globular proteins. A more complete solution of the problem of finding relations between site properties and environment demands further analysis, involving the conformation and physicochemical properties of sites and their environment. Application of methods of molecular dynamics and conformation analysis is promising for further study of this problem.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Bagley S.C., Altman R.B. (1995). Characterizing the microenvironment surrounding protein sites. Protein Sci. 4, 622-635.
2. 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. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. J. Mol. Biol. 112, 535-542.
3. Chothia C. (1976). The nature of the accessible and buried surfaces in proteins. J. Mol. Biol. 105, 1-14.
4. Davies D.R., Cohen G.H. (1996). Interactions of protein antigens with antibodies. Proc. Natl Acad. Sci. USA. 93, 7-12.
5. Ivanisenko V.A., Eroshkin A.M. (1997). Search for sites containing functionally important substitutions in series of related or mutant proteins. Mol. Biol. (Mosk.). 31, 880-887.
6. Ivanisenko V.A., Debelov V.A., Matsokin A.M., Pintus S.S., Grigorovich D.A., Kolchanov N.A. (2002). PDBSiteScan: a tool for search for best-matching superposition over the PDBSite database. This volume.
7. Kabsch W., Sander C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577-2637.
8. Sekharuda Y.C., Sundaralingam M. (1988) A structure-function relationship for the calcium affinities of regulatory proteins containing "EF-hand" pairs. Proteins Eng. 2, 139-146.

PDBSITESCAN: A TOOL FOR SEARCH FOR THE BEST-MATCHING SUPERPOSITION IN THE DATABASE PDBSITE

[1,2] Ivanisenko V.A., [3] Debelov V.A., [4] Pintus S.S., [3] Matsokin A.M., [3] Nikolaev S.V., [1] Grigorovich D.A., [1] Kolchanov N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
[4] Novosibirsk State University, Novosibirsk, Russia
e-mail: salix@bionet.nsc.ru
*Corresponding author

Key words: *PDBSite, active sites, functional sites, 3D structure, best matching, superposition*

Summary

Motivation: Recognition of functional sites in proteins is one of important approaches to the understanding of their functions. The 3D structures of many active sites of proteins, binding sites for various ligands, and biochemical modification sites are stored in the database PDBSite (Ivanisenko et al., 2002). Detection of domains with close structural similarity to sites from PDBSite in the 3D structures of proteins can be very important for recognition of new functions of these proteins.

Results: We developed the program PDBSiteScan, which automatically performs the best superposition of sites from PDBSite with the 3D structure of a protein under study. This allows detection of potential active sites in the 3D structure and gives their structural alignment with known functional sites stored in PDBSite.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html.

Introduction

Proteins containing fragments of spatial structure similar to functional sites of other proteins often have also similar biological properties (Branden, Tooze, 1991). The database PDBSite stores the data on more than 4000 various protein sites, and this information is rapidly supplemented with new data (Ivanisenko et al., 2002). The database contains data on ligand-binding sites, active centers of various enzymes, sites of biochemical modification (phosphorylation, glycosylation, etc.), and so on. Identification of sites with close structural resemblance to sites from PDBSite in the spatial structures of proteins would allow recognition of new functions of these proteins.

The goal of this study was to develop software for construction of the best superposition of sites from PDBSite and the spatial structure of a protein to be analyzed.

Methods for comparison of 3D protein structures were discussed in many papers. The most important of them were reviewed in (Gibrat et al., 1996; Eidhammer et al., 2000). The method for structural comparison required for our purposes should have had the following features. First, it should be rapid enough to compare the protein analyzed with all the sites of PDBSite during a reasonable time period. Second, the comparison should be sought only for a complete site. We suggested that the presence of a similarity to a mere fragment of a site would reduce its functional significance. Third, the method should not only detect a structural similarity but also check amino acid matches between the site and the corresponding region of the protein. In our opinion, the absence of amino acid sequence conservation between the site and the protein can also reduce the functional significance of the similarity. For proper consideration of amino acid diversity in specified positions of a site, the so-called patterns or fingerprints, should be constructed (Bork, Koonin, 1996). The solution of this task is projected for the future.

Most sites in PDBSite contain not more than ten residues, three to four on the average. This allows simple approaches to the search for structural similarity between sites. We developed an algorithm based on exhaustion of all the possible combinations of protein positions to be compared with a site. Limitations on amino acid composition allowed development of a rapid method for search for structural similarity between sites.

Methods and algorithms

The best superposition of two sites represented by a 3D point assemblage

The problem of "best superposition" of two ordered sets $P = \{P_1, P_2, ..., P_N\}$ and $Q = \{Q_1, Q_2, ..., Q_N\}$ of points from R^3 implies determination of a rigid body conversion (x_0, T): $R^3 \to R^3$, where x^0 is a slip vector and T is the matrix of a rotation minimizing the functional

$$F(x^0, T) = N^{-1} \sum_{i=1}^{N} | T(P_i - x^0) - Q_i |^2 = \inf_{x^0, T} F \cdot$$

Its solution is determined by conversions shifting the centers of mass of the sets to the origin of coordinates and determining the rotation matrix (three rotation angles) from the condition

$$F(T(\varphi_1, \varphi_2, \varphi_3)) = \inf_T \left\{ 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot (T P_i, Q_i)}{| P_i |^2 + | Q_i |^2} \right\}$$

by the gradient method.

The initial approximation for the gradient method is determined as the rotation matrix of the solution of the problem of superposition of two three-dimensional triangles whose apices are centers of mass of three nonintersecting subsets of the corresponding initial set. This task is solved by the direct method, and the solution provides the lower estimate for the desired minimum.

Search for structural similarity

All fragments of protein (B) structurally similar to a site under consideration (A) should be found. In our approach, we consider only atoms $N, C\alpha, C$ of the residues. As in (Pennec, Ayache, 1998), each three atoms are regarded as a triangle that is uniquely determined by the location of the atom $C\alpha$ and the reference point of the local coordinate system.

Thus, we have a sequence of spatial triangles $B = \{T_i\}_{i=1}^{n}$ with apices $(N_i, C\alpha_i, C_i)$ and such triangles that for any triangle pair T_i and T_j there exist a "rigid-body" transformation (consisting of a slip and turn), converting one triangle into the other. We should find a substructure B, similar to a sequence of triangles $A = \{T_j(A)\}_{j=1}^{m}$ defined in the same way but short. All substructures $B_{i1,...,1m} \subset B$ of the length m are compared with A, actually superposing A to $B_{i1,...,1m}$. In doing so, the slip vector and the rotation matrix of the rigid-body transformation are determined according to the first triangles of the sequences to be compared. The distance between the transformed sequences is calculated to estimate their similarity.

Results and discussion

We have developed a program for structural alignment of sites from the base PDBSite with the 3D structure of a user-defined protein. The program selects all variants for which the root mean square deviation (RMSD) for atoms $N, C\alpha, C$ does not exceed a level specified by the user. Match of amino acid types for aligned residue pairs is a necessary condition in this implementation of the method.

Proteins used for testing the system PDBSiteScan

PDB ID	Chain	Site residue		
1BIF		256H	325E	390H
1ELV	A	460H	514D	617S
1BN7	A	117D	141E	283H
1AUO	A	114S	168D	199H
1BQY	A	57H	102D	195S
3PVI	A	58D	68E	70K
1CVL		87S	263D	285H
1QLW	A	206S	230E	298H
1B6G		124D	289H	260D
1Ea5	A	200S	327E	440H

Let us consider the application of the approach by the example of recognition of catalytic sites in the family HYDROLASE. A sample of nonhomologous proteins of this family was constructed (similarity less than 40%). Similarities were estimated with the use of data available from the Website (http://www.fccc.edu/research/labs/dunbrack/culledpdb.html). Only proteins with known tertiary structures and known catalytic sites were considered. These proteins and the corresponding sites are listed in the Table.

For analysis of each protein, the catalytic site of this protein was excluded from PDBSite. Figure 1 shows the plots of type I and type II errors for various threshold levels of RMSD.

Type I errors were calculated as proportions of unidentified catalytic sites. Type II errors were estimated by two methods. In the first method (curve **a**), false predictions of catalytically active sites were regarded as errors. In the second method (curve **b**), predictions of new sites of other types, whose presence and location in proteins examined was not strictly defined, were also regarded as errors, in addition to those considered in the first method. Figure 1 shows that catalytically active sites are recognized with a good accuracy. Figure 2 depicts the result of recognition of a catalytic center in the protein 1ELV with the use of PDBSiteScan. The operation of the system revealed a high structural similarity of the catalytic center of this protein to a number of catalytic sites of the base PDBSite. The spatial superposition of one of them and the catalytic center of 1ELV is shown.

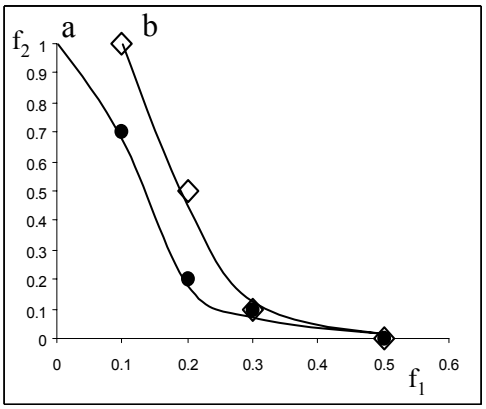


Fig. 1. Relationship between type I and type II errors for recognition of catalytic centers in the HYDROLASE sample. Abscissa (f_1): type I errors; ordinate (f_2): type II errors. Curve (**a**) shows the dependence for a case when false predictions of catalytically active sites were taken as type II errors. Curve (**b**) corresponds to false predictions of catalytically active sites plus predictions of new sites of other types, whose presence and location in proteins examined is not strictly defined.

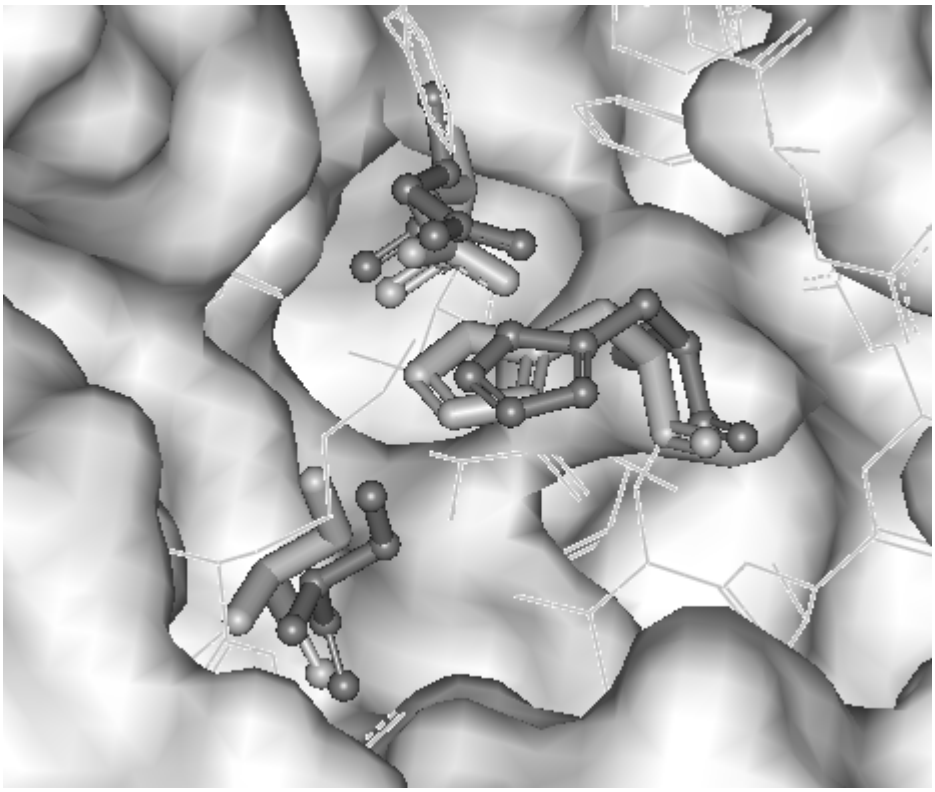


Fig. 2. Spatially superposed residues of the catalytic center of the 1ELV protein and a catalytic site from the database PDBSite (ID 1BQYB), demonstrating the result of PDBSiteScan operation. The residues of the catalytic center of 1ELV are depicted with the use of the "Stick" model, and the residues of the 1BQYB site are shown as "Stick and Ball". The image was made with the use of ViewerLite.

New potential sites were found for each protein of the tested set. Most of them are NAD-binding sites or binding sites of various metal ions. Thus, the system PDBSiteScan is an efficient tool for investigating the functions of proteins and their biologically significant sites.

Conclusion

In the future, modification of the algorithm for accelerating the comparison of 3D structures is planned. It will include preliminary reworking of the database of atom coordinates, use of geometrical hashing, etc. We intend to increase the accuracy of site recognition by taking into account their spatial environment. Moreover, development of a database of 3D patterns of biologically significant sites on the base of PDBSite is planned.

Several protein families will be investigated to reveal new sites in their 3D structure.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Bork P., Koonin E.V. (1996). Protein sequence motifs. Curr. Opin. Struct. Biol. 6, 366-376.
2. Branden C., Tooze J. (1991). Introduction to Protein Structure. Garland Publishing, New York, London.
3. Eidhammer I., Jonassen I., Taylor W.R. (2000). Structure comparison and structure patterns. J. Comput. Biol. 7, 685-716.
4. http://www.fccc.edu/research/labs/dunbrack/culledpdb.html
5. Ivanisenko V.A., Grigorovich D.A., Kolchanov N.A. (2002). PDBSite: database on protein active sites and their environment. This volume.
6. Pennec X., Ayache N. (1998). A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics. 14, 516-522.
7. Smits B. (1998). Efficiency issues for ray tracing. J. Graphics Tools. 3, 1-14.

COMBINING BIOINFORMATICS AND STRUCTURAL METHODS FOR ANALYSIS OF KEY FUNCTIONAL RESIDUES IN DNA REPAIR ENZYMES

[1,] Zharkov D.O., [2] Grollman A.P*

[1] Institute of Bioorganic Chemistry, SB RAS, Novosibirsk, Russia
[2] State University of New York at Stony Brook, Stony Brook, NY, USA
e-mail: dzharkov@niboch.nsc.ru
[*]Corresponding author

Key words: *DNA repair, glycosylase, multiple alignment, clusters of orthologous groups, structure, substrate specificity, reaction mechanism*

Resume

Motivation: An essential function of DNA glycosylases is the recognition and excision of damaged bases in DNA, thereby preserving genomic integrity. Features of lesion recognition are only partially revealed by structural analysis of the enzymes. The functional role of key enzyme residues can be predicted by combining structural data with analysis of amino acid conservation.

Results: The following postulate underlies our approach: if a family or superfamily can be broken into subgroups with different substrate specificities, residues highly conserved between these subgroups represent those important for enzyme catalysis and structure maintenance while residues highly conserved within a subgroup but not between the subgroups represent residues important for substrate specificity. We apply the analysis of multiply aligned sequences (AMAS) algorithm in the clusters of orthologous groups (COG) for quantitative treatment of similarity and dissimilarity in the Nth family of DNA glycosylases. Mapping highly similar and dissimilar regions on three-dimensional protein structures provide a starting point in planning site-directed mutagenesis to elucidate the role of residues important for group-specific functions in DNA repair proteins. The method is universally applicable for families of proteins with diverse functions.

Availability: The primary data on which the analysis is based can be accessed at http://www.pharm.sunysb.edu/lcb/FRBM2002/; software and databases required for implementation of this analysis were collected from publicly available sources:

AMAS, http://barton.ebi.ac.uk/servers/amas_server.html

COG, http://www.ncbi.nlm.nih.gov/COG/

Introduction

DNA repair is critical for maintaining genome stability. Base excision repair, one of the processes acting to prevent cytotoxic or mutagenic effects of DNA damage, involves multiple enzymatic species. During its first stage, damaged bases are excised from DNA by DNA glycosylases. These enzymes vary in specificity: for example, *E. coli* formamidopyrimidine-DNA glycosylase (Fpg) recognizes oxidatively damaged purines; endonuclease III (Nth), oxidatively damaged pyrimidines; while 3-methylpurine-DNA glycosylase (AlkA), ring-alkylated purines.

DNA glycosylases may be divided into several families based on their primary sequences (Eisen, Hanawalt, 1999). For this analysis, which is concerned with functional rather than evolutionary relationships, we define a "family" based exclusively on sequence homology and not phylogeny, including both orthologous and paralogous proteins. The largest family of DNA glycosylases is the Nth family including the prototype Nth and the mismatch adenine DNA glycosylase MutY (Thayer et al., 1995). Several conserved structural motifs are common to these enzymes and other DNA glycosylases (such as AlkA), defining the Nth superfamily (Thayer et al., 1995). Another well-defined group of DNA glycosylases is the Fpg family, comprising Fpg and endonuclease VIII (Nei; Eisen, Hanawalt, 1999).

The structures of several DNA glycosylases have been solved, in some cases, complexed with DNA. Although functionally important residues can be identified from these structures, a wealth of biochemical data cannot be interpreted by structural analysis alone. Analysis of protein conservation, coupled with structural information, provides a valuable tool for predicting functional roles for critical residues and designing site-directed mutants of DNA glycosylases. The existence of enzymes with different specificities within the same family of DNA glycosylases allows searching for protein residues important for catalysis or specificity. To provide proof of principle, we present detailed analysis of the Nth family of DNA glycosylases.

Methods

If a group of related enzymes may be broken into two subgroups with related but different specific functions, then: (i) residues highly conserved between these subgroups are important for general catalytic mechanism and structure

maintenance, and (ii) residues highly conserved within but not between subgroups ("dissimilar" residues) are important for a subgroup-specific function. To quantitatively account for similarity of residues, AMAS algorithm (Livingstone, Barton, 1993) was used. Upon comparing physico-chemical properties of amino acids in a given position of a multiple alignment produced by ClustalW, based on the standard Taylor set of properties (Taylor, 1986; cysteines reduced, 10% minimum residue occupancy), a "conservation number" C_n was assigned to each position. Amino acids were identified as similar or dissimilar following the hierarchical set of rules described below.

To sort enzymes of a family into functional groups, their substrate specificity must be known. Such data are usually available for only a handful of members of the family. The easiest way to circumvent this problem is to start with clearly functionally defined paralogs and to split the family into non-overlapping groups homologous to each of these on the basis of elements known to be critical for the enzyme's function. For Nth and MutY, qualification was based on the presence of four cysteines in the iron-sulfur cluster; the sequences containing lysine at position 120 (*E. coli* numeration) and no C-terminal domain were considered Nth, the sequences containing no lysine and a C-terminal domain were considered MutY.

The current distribution of available sequences is heavily skewed in favor of certain phyla of eubacteria, which might lead to an artificial increase in conservation numbers if all sequences in databases are included in the analysis. A subset of sequences evenly distributed through major taxonomic branches may be chosen to better reflect the functional significance of amino acid similarity. The Cluster of Orthologous Groups tool (Tatusov et al., 1997) was used for this purpose. COGs currently pool data from 44 complete genomes, representing 30 different phylogenetic lineages. Nth COG0177 comprises 47 bacterial proteins of 25 phylogenetic groups and MutY COG1194 consists of 28 proteins from 17 groups. Following qualification, 41 proteins of 24 phylogenetic lineages of the Nth COG and 26 proteins of 16 lineages of the MutY COG were included in the analysis.

A hierarchical set of rules to avoid ambiguities in similarity/dissimilarity assignment was designed. Different positions were classified into the following categories: "Identity between all subgroups", "Identity within one subgroup", "Conservation between all subgroups", "Difference between subgroup pairs", "Conserved within one subgroup" and "Unconserved within one subgroup". We considered (former overruling latter):

1. residues identified as "identity between all subgroups" as similar in both subgroups;

2. residues identified as "difference between subgroup pairs" as dissimilar in both subgroups;

3. residues identified as "conservation between all subgroups" as similar (if C_n>8) in both subgroups;

4. residues identified as "identity within one subgroup" as dissimilar in this subgroup;

5. residues identified as "conserved within one subgroup" as dissimilar (if C_n>8) in this subgroup.

After highly conserved and dissimilar residues have been identified, they were mapped on known structures of proteins under comparison. Structures of *E. coli* Nth (Thayer et al., 1995; 2ABK) and the p25 domain of *E. coli* MutY (Guan et al., 1998; 1MUY) were used for mapping.

Results and Discussion

Nth catalyzes removal of oxidized pyrimidines and employs Schiff base chemistry for nucleophilic attack by Lys-120 at C1' of the damaged nucleotide. MutY excises mispaired adenine through attack by an activated water molecule. MutY consists of two domains readily separated by proteolysis. The N-terminal p25 domain retains catalytic activity and is homologous to Nth. Both Nth and MutY contain an iron-sulfur cluster that is required for activity but not directly involved in catalysis. The structures of Nth and p25 of MutY show close similarity in overall folding, but structural information on their interactions with DNA is not available.

In Nth, 12% of the 211 amino acid residues are similar and 9% dissimilar; in MutY, these values are 12% and 15%, respectively. Most similar and dissimilar residues face the DNA-binding cleft. The only region of similarity fully buried inside the protein globule is composed of hydrophobic residues in α-helices 2, 4 and 5 in the six-helix barrel domain, probably crucial for interhelical packing. Other similar residues may be broken in four patches. The largest of these (S1, Fig.) is an iron-sulfur cluster and its surroundings. Two others (S2 and S3) are centered around the catalytic dyad of the Nth superfamily, 138 and 120, almost coinciding with the conserved G/P...D loop motif and the helix-hairpin-helix motif, respectively (Thayer et al., 1995). Obviously, these elements are absolutely required for correct positioning of key catalytic amino acid residues in both enzymes. The smallest patch of similarity (S4) involves two residues at the six-helix barrel domain edge of the DNA-binding cleft and may participate in DNA binding.

Dissimilar residues in Nth and MutY may play a crucial role as determinants of the markedly different substrate specificity of these two enzymes. Most dissimilar residues in Nth are concentrated on the DNA-binding face of the protein while MutY has a patch of dissimilar residues on the opposite face (see below). As expected from the reaction mechanism, position 120 is dissimilar in Nth, but, unexpectedly from structural and biochemical information for the *E. coli* enzyme, it is not dissimilar in MutY. This position in the MutY family is occupied by either Ser or Tyr, residues of quite different properties. The residues spatially close to position 120 form distinct dissimilarity patches (D1) inside deep pockets of these enzymes where the everted nucleotide presumably is bound. One of these in MutY, E37, was identified by structural and mutagenic analysis as a crucial residue for recognition of adenine (Guan et al., 1998).

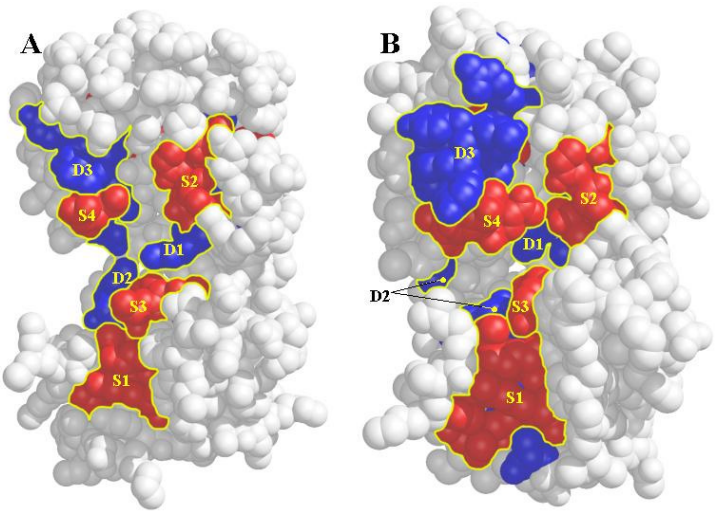


Fig. Structures of *E. coli* Nth (A) and the catalytic domain of *E. coli* MutY (B). Residues identified as similar are colored red, dissimilar residues are blue, all others are shaded grey. The margins of certain clusters of similarity or dissimilarity are traced with yellow lines. Individual residues and clusters are labeled (see main text for details).

A number of dissimilar residues are present at the edges of the DNA-binding groove (D2) in both enzymes, forming "lips" around the "mouth" of the groove. For MutY, V45 and N140 have been implicated in the mechanism of base excision (Guan et al., 1998). DNA glycosylases generally insert several amino acid residues into the DNA double helix, assisting in eversion of the damaged nucleotide and recognition of the opposing base. The residues listed above are good candidates for performing these functions.

The largest region of dissimilarity (D3) is found in the six-helix barrel domain. In Nth, it surfaces on the DNA-binding face of the enzyme and penetrates deep into the hydrophobic core of the domain while, in MutY, it runs through the entire domain, from the DNA-binding face of the enzyme to the opposite face. This region on the surface of the DNA-binding face of MutY is rich in hydrophobic residues exposed to the solvent and projects approximately in the same direction as the C-terminal domain of MutY (absent in the core domain structure). These residues may be involved in interactions with the C-terminal domain, which is proposed, based on NMR data and molecular modeling, to fold back on the barrel domain of MutY (Volk et al., 2000). Buried residues of this region may play a role in overall shape maintenence, accounting for the large observed difference in the size of the DNA-binding cleft of Nth and MutY.

We also performed similar analysis for the enzymes of Fpg family, and for enzymes of Nth and AlkA subgroups of the Nth superfamily. In both cases, the subgroup function-specific residues were identified. Site-directed mutagenesis was conducted to confirm the functional role of the Nei R212 residue predicted to participate in substrate recognition. The results will be presented in full in a less restricting format.

Acknowledgements

This study was supported in part by the grant 02-04049605 from the Russian Foundation for Basic Research. The authors are grateful to J. Sussman for helpful discussions.

References

1. Eisen J.A., Hanawalt P.C. (1999) A phylogenomic study of DNA repair genes, proteins, and processes. Mutat. Res. 435, 171-213.
2. Guan Y., Manuel R.C., Arvai A.S., Parikh S.S., Mol C.D., Miller J.H., Lloyd R.S., Tainer J.A. (1998) MutY catalytic core, mutant and bound adenine structures define specificity for DNA repair enzyme superfamily. Nat. Struct. Biol. 5, 1058-1064.
3. Livingstone,C.D., Barton G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. Comput. Appl. Biosci. 9, 745-756.
4. Tatusov R.L., Koonin E.V., Lipman D.J. (1997) A genomic perspective on protein families. Science. 278, 631-637.
5. Taylor W.R. (1986) Classification of amino acid conservation. J. Theor. Biol. 119, 205-218
6. Thayer M.M., Ahern H., Xing D., Cunningham R.P., Tainer J.A. (1995) Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. EMBO J. 14, 4108-4120.
7. Volk D.E., House P.G., Thiviyanathan V., Luxon B.A., Zhang S., Lloyd R.S., Gorenstein D.G. (2000) Structural similarities between MutT and the C-terminal domain of MutY. Biochemistry. 39, 7331-7336.

CLASSIFICATION OF LOCAL SPATIAL ENVIRONMENT OF AMINO ACID RESIDUES BY PHYSICOCHEMICAL CHARACTERISTICS: ANALYSIS OF TRANSCRIPTION FACTOR DNA-BINDING DOMAINS

**** Afonnikov D.A., Nikolaev S.V., Ivanisenko V.A.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: ada@bionet.nsc.ru

Key words*: protein structure, amino acid physicochemical properties, protein microenvironment, classification, hierarchical clustering, DNA binding*

Resume

Motivation: Interaction of an amino acid residue with its local spatial environment contributes significantly to the pattern of mutation fixing. Heterogeneity of the local environment brings forth the differences in evolutionary modes of amino acid residues, even within the same protein. This arises the problem of estimating the heterogeneity in local environment of individual residues in proteins.

Results: A method for estimating the heterogeneity in local environment of individual residues in proteins utilizing the cluster analysis of averaged physicochemical characteristics of the environment in question is proposed in this work. The method was applied to analysis of local environments of amino acid sites within DNA-binding domains of transcription factors. Nine classes of local environment were detected, and their interrelations with the structural characteristics of these sites were analyzed. The similarity of local environments of various types of amino acids was analyzed.

Introduction

The local environment of an amino acid residue in many ways determines the pattern of its substitution (Grantham, 1974). The local environment of residues in protein structures is heterogeneous, namely, the environment is predominantly hydrophobic inside the globule and polar on its surface (Chothia, 1984). Different protein regions may also have a different secondary structure or may bind a ligand molecule. This heterogeneity results in different patterns of amino acid substitutions. The diversity of these patterns is mainly taken into account through constructing matrices of the amino acid substitutions specific of certain local environment classes. As a rule, the classes of environments are considered with reference to the secondary structures of amino acid sites and their availability to solvent (Wako, Blundell, 1994a, b; Koshi, Goldstein, 1995). However, these classifications are based on the known structural properties of amino acid sides of proteins and are introduced by researches *a priori.*

In this work, we are classifying the local spatial environment of the residues utilizing averaged values of physicochemical properties without an explicit use of any structural information. We studied DNA-binding domains of transcription factors. The classification thus obtained is used to analyze the dependences on the type of secondary structure and exposure of amino acid site. The results obtained demonstrate that the classification basing on physicochemical properties allows the environments to be grouped according to two characteristics—availability to solvent and types of secondary structures. Similarity between amino acids with reference to their occurrence rates in different classes of local environments was analyzed.

Materials and Methods

Main stages of analysis. The analysis involved the following stages: (1) determining local spatial environment of residues; (2) calculating physicochemical characteristics of the local environment of individual residues; (3) clustering the local environment with respect to their physicochemical properties; (4) analyzing the resulting classes of physicochemical characteristics of the environment and their interrelations with the structural properties of residues; and (5) analyzing the similarity of spatial environments for various amino acid types.

Data. Spatial structures of the transcription factors displaying a degree of sequence similarity not exceeding 40% were used for the analysis. These structures included (chain names are indicated in parenthesis): 1AIS(B), 1BH9(A,B), 1BM8, 1BOR, 1BVO(A), 1C7U(A),1CF7(A,B), 1CI6(A), 1CQT(I), 1D8J(A), 1DH3(A), 1DL6(A), 1DP7(P), 1ENW(A), 1EO0(A), 1EQF(A), 1EXE(A), 1F3U(A,B), 1F4S(P), 1F62(A), 1G2Y(A), 1GD2(E), 1HKS, 1HLO(A), 1I27(A), 1I4W(A), 1JFI(A,B), 1K99(A), 1MNM(A,C), 1NCS, 1PUE(E), 1QQH(A), 1SKN(P), 1SP1, 1TBA(A,B), 1TF3(A), 1TFI, 1YTF(B,C), and 3HSF. The total number N of the residues analyzed amounted to 4240.

Structural properties of residues. We used two parameters, calculated by the program DSSP (Kabsch, Sander, 1983), as structural properties, namely, secondary structure of residues and the surface area accessible to water. These characteristics were calculated for each protein chain separately from others.

Determination of local spatial environment of residues. It was assumed that the local spatial environment of the *i*th (central) residue is formed by the residues whose Cα atoms were located at a distance not exceeding 7 Å from the Cα atom of the *i*th residue. No exclusions were made for the immediate neighbor residues. We considered that the neighbors immediate in the primary structure also contribute to the local environment of a residue in question.

Physicochemical characteristics of the local environment of individual residues. Values of five amino acid properties, taken from (Bogardt et al., 1980)—volume, polarity, isoelectric point, hydrophobicity, and the surface area accessible to water—were used while describing the physicochemical characteristics. The spatial environment of the *i*th residue was characterized with the vector $f_i = \{f_k, k = 1,\ldots 5\}$; each component of the vector corresponded to the value of *k*th amino acid property averaged over the resides forming the local spatial environment.

Classification of physicochemical properties of local environment. To classify the local spatial environment of individual residues according to their physicochemical characteristics, we used the hierarchical cluster analysis (Sneath, Sokal, 1973). The squared Euclidian distance $d_{ij} = [\sum_{k=1}^{5} (f_{ik} - f_{jk})^2]$ was used as a measure of the distance between characteristics of the environment for a pair of residues *i, j.* Unweighted paired grouping method with arithmetic mean (UPGMA) was used for constructing the similarity tree.

Comparison of the observed and expected occurrence numbers of secondary structure types. Upon dividing the local environments into *C* classes, we calculated the observed numbers of occurrence n_{sc} of various secondary structure types of *s* residues (according to the DSSP classification) in the class *c* environment. The occurrence rates of the secondary structure class were estimated as $p_s = 1/N \sum_{a=1}^{20} n_{sc}$; the occurrence rates of the local environment class, as $p_c = 1/N \sum_{c=1}^{C} n_{ac}$; their difference $dn_{sc} = n_{sc} - Np_s p_c$ was used to compare the deviation of the observed numbers of occurrence from the expected numbers.

Analysis of the similarities between environments of different amino acid types. The difference between the number of occurrence n_{ac} of type *a* amino acid in the class *c* spatial environment and its expected value $dn_{ac} = n_{ac} - Np_a p_c$ was calculated analogously. At the final stage, we analyzed similarities of amino acids according to their occurrence rates in different classes of local environment. For this purpose, we used the hierarchical clustering by UPGMA. As a measure of the distance between amino acids of types *A,B,* $d = 1 - r_{ab}$ was used, where r_{ab}, the correlation coefficient between the values of $dn_{ac}(A)$ and $dn_{ac}(B)$. This measure was selected, as in this particular case we were interested in the relative occurrences of individual amino acid types in different local environment classes.

Results and Discussion

The analysis allowed us to divide the local environments of all the residues into nine classes. Profiles of the averaged physicochemical characteristics for individual local environment classes are shown in Fig. 1; the number of residues in the class, mean accessibility of amino acid residues to the solvent (according to DSSP), difference in the observed and expected numbers of occurrence of the secondary structure classes are listed in Table.

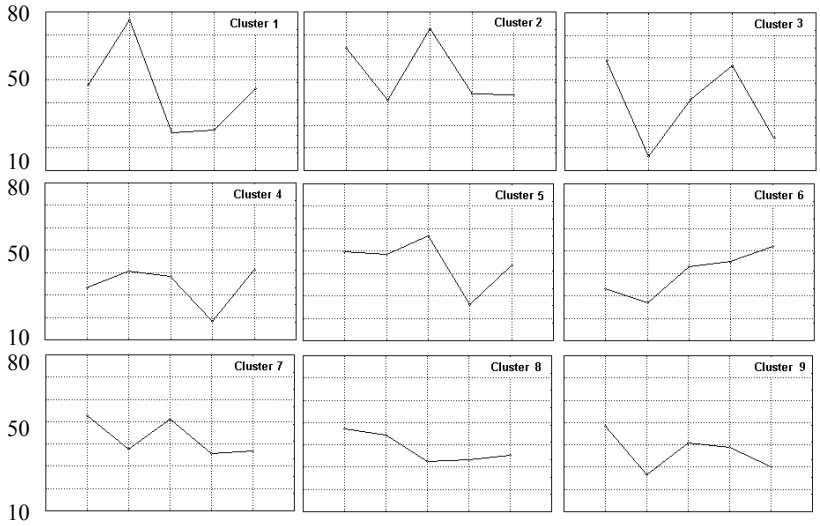


Fig. 1. Profiles of the average values of the five residue properties in the nine classes of spatial environment (all the plots have the same scale): V, volume; P, polarity; pI, isoelectric point; HP, hydrophobicity; and SA, the surface area accessible to water (the X axis).

Table. Distribution the structural properties of the central residue in the classes of local environment.

Class index	1	2	3	4	5	6	7	8	9
Class Size[a]	14	37	147	417	373	7	693	887	1665
Mean Acc.[b]	135.8	131.6	60.9	99.0	97.3	116.1	71.2	75.4	62.0
dn_{sc}(X)[c]	**3.0**	**13.0**	3.4	**49.3**	11.8	**4.5**	–23.0	–13.8	–48.1
dn_{sc}(H)[c]	–3.9	–7.6	–26.9	–83.5	**21.1**	–2.9	**50.4**	**50.8**	2.4
dn_{sc}(T)[c]	1.3	–1.4	–3.4	30.5	–2.2	–0.8	1.8	11.8	–37.5
dn_{sc}(S)[c]	1.8	0.7	–1.9	34.3	11.1	0.4	–23.0	–0.0	–23.5
dn_{sc}(G)[c]	–0.3	–0.8	–2.2	–2.0	–2.1	–0.2	5.0	4.8	–2.1
dn_{sc}(E)[c]	–1.8	–3.9	**30.7**	–26.8	–38.0	–0.9	–11.0	–51.5	**103.3**
dn_{sc}(B)[c]	–0.1	–0. 2	0.3	–1.9	–1.8	–0.0	–0.1	–2.0	5.5

[a]Number of central residues with the local environment belonging to the environment class indicated.

[b]Mean accessibility to solvent (Kabsch and Sander, 1983) calculated for the central residues belonging to the environment class indicated.

[c]dn_{sc}, values of secondary structure types and the environment classes; designations of secondary structure types (Kabsch and Sander, 1983): H, α-helix; B, residue in isolated β-bridge; E, extended strand involved in β-ladder; G, 3-helix (3/10 helix); T, hydrogen bonded turn; S, bend; X, nonstructured region (the maximal values within each class are bold-faced).

As is evident from Fig. 1, the environment classes may be characterized in the following way according to their physicochemical properties. The environment ascribed to the first class is formed by the residues with a high polarity and accessibility to water and charged negatively (low values of isoelectric point). Typical of the second class is a large residue volume and positive charge; of the third, an increased hydrophobicity and large volume; of the fourth, a small volume and low hydrophobicity; and of the fifth class, a small volume and relatively high polarity. The environments of classes 7–9 display predominantly a nonpolar character.

Interestingly, the mean accessibility to water of the central residue complies well with the data on physicochemical properties (Table). The environments of classes 1, 2, and 6 correspond to the residues displaying a high accessibility to solvent (over 110). The accessibility to water of the residues from classes 4 and 5 is close to 100 and may be characterized as moderate. The accessibility of classes 7 and 8 is low (about 70), while classes 3 and 9 exhibit a very low accessibility to solvent (~60). The overall trend corresponds to a decrease in the mean polarity of the environment.

Another interesting feature discovered is the interrelation between the secondary structure type of the central residue and the class of its environment. Characteristic of the central residue belonging to classes 1, 2, and 6 is the prevalence of nonstructured conformation of this residue. Most likely, such residues are located within exposed loops (interestingly, their fraction is relatively low). Classes 4 and 5 display a moderate accessibility of the central residue to solvent; however, they differ in the secondary structure of this residue: predominant in class 4 are nonstructured conformations; in class 5, α-helix. The classes with a low degree of exposure of the central residue (7 and 8) correspond mainly to the α-helix residues; the classes with a very low degree of exposure (3 and 9), to β-structures. The dependence observed may result from the fact that we included the immediate neighbor residues into the local environment.

Thus, this classification reflects in general the clustering of amino acid sites according to the degree of their exposure (displaying the general trend from high to low accessibility to solvent) and types of secondary structure (α, β, and nonstructured).

Clustering of amino acids basing on their spatial environment is shown in Fig. 2. This pattern displays certain similarity to the traditional classification of amino acids. For example, the amino acid groups {K,R,H} and {F,Y,M,I} cluster together, complying with their physicochemical similarity. However, there are some distinctions from the traditional classifications. For example, alanine clusters with positively charged K, R, and H. It was unexpected to find similarity of the groups {N,V} and {D,W,T} (however, typical of the latter group is polar atoms in their side chains). We assume that these differences may be related to the specific features of the structure and function of transcription factor DNA-binding domains.

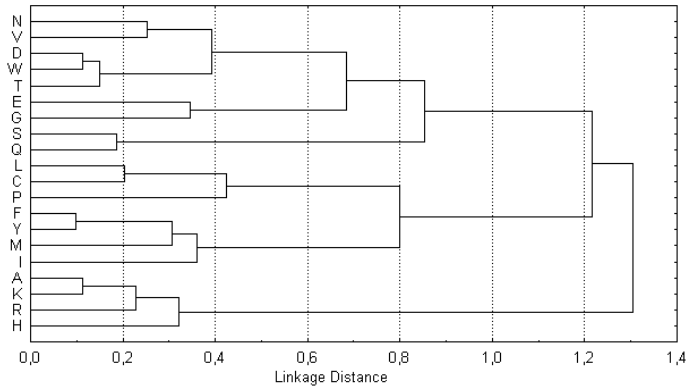


Fig. 2. The similarity tree for 20 types of amino acids constructed according to their occurrence rate in different classes of local environment.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Bogardt R.A.Jr, Jones B.N., Dwulet F.E., Garner W.H., Lehman L.D., Gurd F.R. (1980). Evolution of the amino acid substitution in the mammalian myoglobin gene. J. Mol. Evol. 15:197-218.
2. Chothia C. (1984). Principles that determine the structure of proteins. Ann. Rev. Biochem. 53:537-572.
3. Grantham R. (1974) Amino acid difference formula to help explain protein evolution. Science. 185:862-864.
4. Kabsch W., Sander C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers. 22:2577-2637.
5. Koshi J.M., Goldstein R.A. (1995). Context-dependent optimal substitution matrices. Protein Eng. 8:641-645.
6. Sneath P.H.A., Sokal R.R. (1973). Numerical Taxonomy. The Principles and Practice of Numerical Classification. San Francisco: W.H. Freeman and Co.
7. Wako H., Blundell T.L. (1994a). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J. Mol. Biol. 238:693-708.
8. Wako H., Blundell T.L. (1994b). Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. II. Secondary structures. J. Mol. Biol. 238:693-708.

THE NUMBERS OF PROTEIN DOMAIN SEQUENCES AND PROTEIN CODING GENES IN THE EVOLVED PROTEOMES

**** Kuznetsov V.A., [1] Pickalov V.V.***

The Laboratory of Integrative and Medical Biophysics, National Institute of Child Health and Human Development, NIH, 13 South Drive, Bethesda, MD, 20892, USA, e-mail: vk28u@nih.gov

*Corresponding author

[1] Institute of Theoretical and Applied Mechanics, SB RAS, Novosibirsk, Russia, e-mail: pickalov@itam.nsc.ru

Key words: *proteome complexity, evolution, number of genes, number of protein domains, Pareto-like distribution*

Resume

Motivation: Obtaining an accurate estimation of the number of protein coding genes and the number of protein domains for different organisms is a daunting challenge, even with high-throughput sequence and expression analysis. Our ability to compare the genomes and proteomes of many species might help us to estimate these fundamental numbers for partially-sequenced genome organisms.

Results: We have analyzed the frequency distributions of the protein domain coding DNA sequences (putative protein domain sequences) by their occurrence values in 70 sample genomes of three domains of life: archaeal, bacterial and eukaryotic organisms. All observed domain occurrence frequency distributions (DOFD) are fitted well by the Pareto-like function those shape systematically depends on the number of domains in the sampled genomes. Such family of the frequency distributions is derived from our random birth-death process model of the evolving proteome. We found the functional relationship between the number of distinct domains, number of evolutionarily conserved genes, and the total number of protein domains. We predict ~41,000 protein-coding genes in the human genome, ~21,000 protein-coding genes in the mouse genome and ~5,500 proteome domains in the entire "proteome world".

Introduction

Our ability to statistically compare the protein domain coding DNA sequences for fully-sequenced genomes of many species might help us estimate two fundamental numbers in biology: the number of genes for partially-sequenced eukaryotic genomes and the number of distinct protein domain coding DNA sequences associated with evolutionarily conserved protein domains as the primary evolving units. These estimates might also lead us to better understand the mechanisms of complexity growth for organisms due to evolution. In particular, previous attempts at estimating the number of human gene loci have been predicated on approaches such as measuring the complexity of cellular RNA, reassociation kinetics, CpG island determination, or assuming that cDNA sequences represent genes or evolutionary 'rules of thumb'. However, reliable estimate of the number of human genes has not been obtained yet (Hogenesch et al., 2001). Recently, Rzhetsky and Gomez have estimated the number of pairs of distinct protein domains for *E. coli* (>4,600) and for yeast (>12,000) genomes. These predictions were based on the assumptions that the number of pairs of protein domains is at least three times larger than the number of genes in the genome, and that the distribution of protein domain sequences in genomes is a "scale-free" distribution associated with a simple power law. However, goodness of fit analysis of the protein domain sequence datasets for many fully-sequenced genomes shows that this model is a poor fit to the data (Kuznetsov, 2002). Moreover, a power law statistical model assumes that the structural complexity of a system is invariant relative to the size of the sample system. Such a model can explain the growth of the system, but cannot explain evolution of the systems complexity.

Results and Discussion

By the working definition, a protein domain is defined as an independent structural unit of proteins which can be found alone or in conjunction with other domains or repeat sequences. Protein domains are evolutionarily related. Even though the structure of a domain is not always known, it is still possible in many cases to define the domain boundaries from the protein-coding DNA sequence alone. In this work, the sequence criteria recognized by InterPro data base (www.ebi.ac.uk/interpro/), was used to define a protein domain.

Based on the assumption that basic evolutionary processes can be fruitfully understood in terms of distinct protein domain coding DNA sequences, we have analyzed the statistical distributions of protein domains by their occurrence values in sample proteomes of three domains of life: archaeal, bacterial and eukaryotic organisms. The 68 fully-sequenced genomes as well as incompletely-sequenced and partially annotated mouse and human genomes represented in the InterPro database (12 March, 2002:www.ebi.ac.uk/interpro/) have been analyzed (see Appendix 1). We found that observed domain

occurrence probability distributions (DOPDs) in all these cases have the following characteristics in common: there are few frequent, and many rare distinct domains. This data fits well by the Generalized Discrete Pareto (GDP) probability function:

$$f(m) = z^{-1}/(m+b)^{k+1}, \quad (1)$$

where the function *f(m)* denotes the probability that a distinct, randomly chosen domain, occurs exactly *m* times in the proteome. The function *f(m)* involves two unknown parameters, *k,* and *b*, where *k*>0, and *b*>-1. The normalization factor *z* is the generalized Riemann Zeta-function value: $z = \sum_{j=1}^{J} 1/(j+b)^{k+1}$. *J* is the occurrence value of most common domain. This empirical parameter is a function of the number of occurrences of protein domains, *M*, in the proteome. *M* reflects the total number of the sequences that are potentially functional and, thus, can characterize the proteome complexity.

Commonly, the observed DOPDs does not show the scale-invariant property: a form of the distribution systematically depends on size of the sampled proteome (Fig. a). Figures b shows that the observed occurrence values for the most frequent domain in a proteome is a function of the total number of the observed protein domains in the proteome, *M*. The regression line *J=0.03M* was fitted to data points for all 70 organisms.

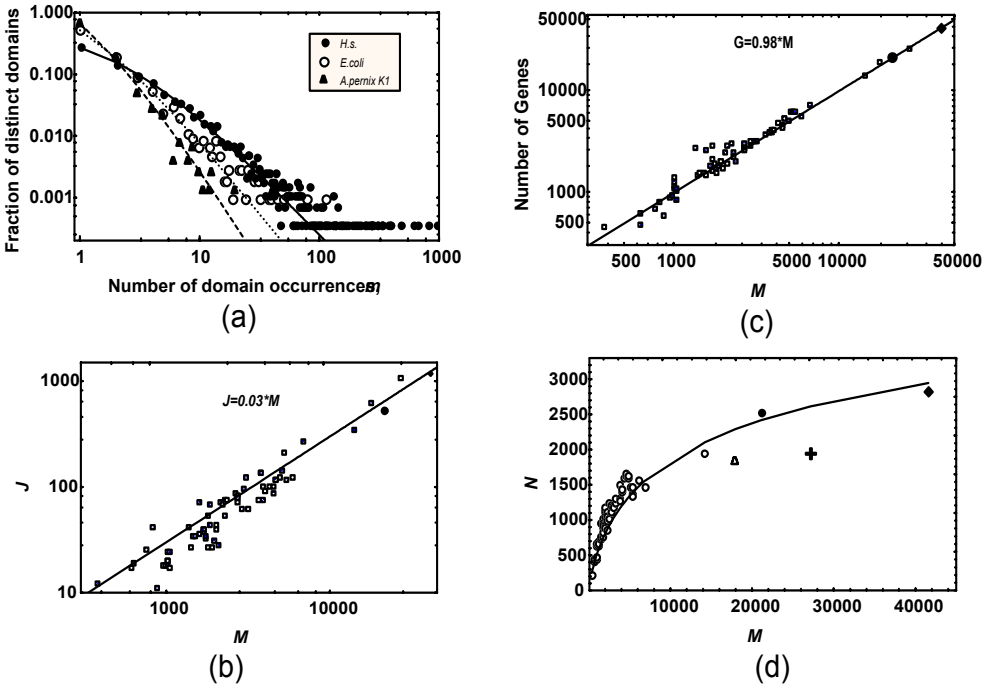


Fig. Characteristics of the evolved proteomes. (**a**). Log-log plot: Fitting of the empirical frequency distributions by the GDP model for human (•: at *k=1.02; b=2.17*), *E. coli* (○ : at *k=1.41; b=0.88*), and *A. pernix K1* (▲: *k=2.07; b=0.77*), respectively). **(b)**. Observed occurrence values for the most frequent domain in the proteome as a function of the number of protein domains in the proteome, *M*. Regression line *J=0.03M* fitted to ○ -data points for 70 organisms. (•,◆): predicted points for mouse and human, respectively (see Appendix 1: List of the organisms ordered by the value *M*). **(c)**. Relationships between the number of evolutionarily conserved protein-coding genes in the genome, *G*, and the total numbers of protein domains in the sample proteomes, *M*, for *H. sapiens*, *M. musculus* and the 68 organisms of fully-sequenced genomes. (○); (•,◆): the predicted points for mouse and human, respectively. The regression line is *G=0.98M* for ○ -data. **(d)**. Fitting of the relationships between the number of distinct protein domains for 68 sample proteomes (except for the data points for *A. taliana* (+) and *C. elegans* (6) by the model Eq.2 at c=0.49+/-0.02; d=4,500+/-300.

Appendix 1.

List of the organisms ordered by the total number of InterPro protein domains in the sample proteome.

Guillardia theta (algal nucleomorph), Ureaplasma parvum, Mycoplasma genitalium, Mycoplasma pneumoniae, Mycoplasma pulmonis, Buchnera aphidicola (subsp. Acyrthosiphon pisum), Chlamydia trachomatis, Chlamydia muridarum, Borrelia burgdorferi, Rickettsia conorii, Chlamydia pneumoniae strain AR39, Chlamydia pneumoniae strain CWL029, Treponema pallidum, Rickettsia prowazekii, Chlamydia pneumoniae strain J138, Aeropyrum pernix K1, Thermoplasma acidophilum, Thermoplasma volcanium, Helicobacter pylori strain 26695, Helicobacter pylori strain J99, Pyrobaculum aerophilum, Methanobacterium thermoautotrophicum, Sulfolobus tokodaii, Pyrococcus horikoshi, Campylobacter jejuni, Mycobacterium leprae, Methanococcus jannaschii, Aquifex aeolicus, Streptococcus pyogenes strain SF370, Pyrococcus abyssi, Neisseria meningitidis strain Z2491 (serogroup A), Neisseria meningitidis strain MC58 (serogroup B), Haemophilus influenzae, Halobacterium sp. NRC-1, Xylella fastidiosa, Thermotoga maritime, Sulfolobus solfataricus, Archaeoglobus fulgidus, Lactococcus lactis (subsp. lactis) strain IL1403, Pasteurella multocida, Clostridium perfringens, Listeria innocua, Staphylococcus aureus strain Mu50, Staphylococcus aureus strain N315, Deinococcus

radiodurans, Listeria monocytogenes, Brucella melitensis, Synechocystis sp. PCC 6803, Caulobacter crescentus, Clostridium acetobutylicum, Yersinia pestis, Bacillus halodurans, Vibrio cholerae, Bacillus subtilis, Salmonella typhi, Salmonella typhimurium, Escherichia coli K-12, Escherichia coli O157:H7 substrain RIMD 0509952, Escherichia coli O157:H7 strain EDL933, Schizosaccharomyces pombe, Anabaena sp. strain PCC 7120, Rhizobium loti, Saccharomyces cerevisiae, Pseudomonas aeruginosa, Rhizobium meliloti, Drosophila melanogaster, Caenorhabditis elegans, Mus musculus, Arabidopsis thaliana, Homo sapiens.

Fig. c shows remarkable linear relationships between the number of evolutionarily conserved protein-coding genes, *G*, and the number of protein domains presented in proteome, *M*. According to the regression line for 68 of the 70 organisms (excluding data points for human and mouse) presented on Figure c, 41,580 protein-coding genes in the human genome and 21,120 protein coding genes in the mouse genome were predicted. Note, that our predictions have been based on the Interpro database, released on March, 12,2002, which used incomplete mouse genome sequences. The latest (May 4, 2002) mouse draft sequence based on whole genome shotgun analysis covering 96% of the mouse euchromatic DNA predicts 22,444 genes (Mouse Genome Assembly v.3; http://www.ensembl.org/Mus_musculus/). The linearity of the graph in figure 1c is surprising. It may be related to small statistical samples of the protein domain sequences. However, even if the average of these proteome represents 60-65% of the total, it seems unlikely that predictions will be changed substantially with adding data. Based on the relationships presented in Figure, we also find that the fraction of unique domains represented by only one entry in the proteome, *f(1)*, becomes smaller when the number of genes in the genome increases (Kuznetsov, 2002). Our analysis suggests that the general trend in the evolution of proteome complexity is characterized by the occasional appearance of a new entry of a domain sequence which already has been encoded in the genome (using a new combination of "old" domains, and less probable the usage of a "new" one).

Based on the relationship between *M* and the number of distinct protein domains, *N*, we can estimate the entire number of distinct protein domains in all proteomes, N_t. Figure d shows that as the number of evolutionarily conserved protein-coding genes increases the number of distinct protein domains, *N*, tends to be limited. The estimate based on curve-fitting of our model (Kuznetsov, 2001)

$$\frac{dN}{dM} = \frac{1+1/d^c}{1+(M/d)^c}\frac{N}{M}, \qquad (2)$$

with *N(1)=1*. Parameters *c* and *d* are positive constants. Eq.2 defines the population "growth" function of the number of distinct protein domains *N(M)*. This function can characterize the proteome complexity growth in the course of evolution (see Fig. d). Eq.2 has an exact solution $N(M) = (M^c \frac{1+1/d^c}{1+(M/d)^c})^{\frac{1+1/d^c}{c}}$ with a limit $N(\infty) = N_t = (1+d^c)^{\frac{1+1/d^c}{c}}$ as $M \to \infty$, where N_t is the total number of protein domains in the "proteome world". This limit was estimated to be 5360±380 distinct protein domains in the proteome world. Interestingly, the *A. thaliana* and *C. elegans* data points on Figure d significantly drop downward from a predicted general trend. We excluded these points from curve-fitting analysis. Hence many genes in genomes of *A. thaliana* and *C. elegans* have been duplicated in the course of evolution, than the negative differences between the best-fit model data point by Eq.(2) and observed data can be used as the estimator of the fraction of duplicated genes for a specific organism.

We simulated the "reverse" process of the protein domain evolution using Monte-Carlo method (i.e. going back one evolutionary step to a simpler proteome). This procedure was based on the random removal of one domain copy from a sample proteome. We observed that the DOFD for the sub-samples of the more complex proteome organism (i.e. human) was transformed to DOFDs for the simpler organisms (i.e. worm, yeast, *E. coli*, consequently) regardless of how many protein domains in comparing samples are not identical. Based on similarity of the DOFDs for random sub-samples from a larger proteome and for any smaller proteomes we suggest that a randomness is likely the predominant process in the evolution of the proteome world.

To better understand the relationship between processes of the proteome evolution and dynamics of the DOFDs, we developed a stochastic model displaying domain occurrence in the proteome during the course of evolution. If time parameter *t* is the continuos parameter and the domain occurrence process has the birth and death properties, than the probability *pm* that a given protein domain has occurred exactly *m* times in the proteome can be described by the differential-difference equations for the birth-death process (Kuznetsov, 2002a), where intensities of the birth and death process are linear functions of the occurrence value m, i.e:

$dp_0(t)/dt = -\lambda_0(t)p_0(t) + \lambda_1 p_1(t)$;

$dp_m/dt = -(\lambda_m + \mu_m)p_m(t) + \lambda_{m-1}p_{m-1}(t) + \mu_{m+1}p_{m+1}(t)$;

$m = 1,2,\ldots$; $\lambda_m = \lambda^*_1 + m\lambda^*_2$; $\mu_m = \mu^*_1 + m\mu^*_2$. Under specification of the parameters of the model, the quasi-steady state solution of the model has the family of the distributions including Pareto, Yule-Simon and Waring-Irwin distributions as the specific cases. The Rzhetsky and Gomez model has the Yule's formula at asymptotic limit. The distribution family of our model combines specific properties: the long left tail and the size (or time) dependence of the distribution shape.

References

1. Hogenesch J.B., Ching K.A. et al. (2001) A comparison of the Celera and Ensembl predicted gene sets reveals little overlap in novel genes. Cell. 24, 413-415.
2. Kuznetsov V.A. (2001) Distribution associated with stochastic processes of gene expression in a single eukaryotic cell. EURASIP J. on Applied Signal Processing. 4, 285-296.
3. Kuznetsov V.A. (2002) Statistics of the numbers of transcripts and protein sequences encoded in the genome. In: Computational and Statistical Methods to Genomics. (W.Zhang, I.Shmulevish, eds.) Kluwer: Boston. 125-171.
4. Kuznetsov V.A. (2002a) A Family of Skewed distributions generated by the stochastic processes of molecular biology systems. Signal Processing (submitted).
5. Rzhetsky A., Gomez S.M. (2001) Birth of scale-free molecular networks and the number of distinct DNA and protein domains per grnome. Bioinformatics. 17, 988-996.

DESIGN OF A KNOTTED CUBIC-LATTICE PROTEIN

[1] Titov I.I., [2] Pal'yanov A.Yu., [1,3] Ivanisenko V.A

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: titov@bionet.nsc.ru
[2] Novosibirsk State University, Novosibirsk, Russia
[3] SRC VB "Vector", Novosibirsk, Russia
*Corresponding author

Key words: *polymer, artificial evolution, protein design, lattice model, knot*

Resume

Motivation: Formation of a knot by a polymer requires overcoming of an entropic barrier complicated by structure compacting due to stability of intermediate states. Consequently, only few proteins containing knots have been discovered; the majority of such proteins are simple and formed of a small piece of chain running through a loop (Taylor, 2000). Designing an artificial model of knotted protein might be helpful in answering the questions on what factors control the topology of protein folding and what is the difference between folding of proteins with complex topology and an ordinary folding.

Results: A 27-mer cubic-lattice protein forming a compact spatial structure with a knot is considered in the framework of a "perturbed homopolymer" model. At the first, thermodynamic, stage of designing, sequences for which a knotted structure is energetically optimal were found. At the second stage, the kinetics of folding into the target structure was optimized. It appeared that not all the native contacts were equivalent in kinetic optimization and only a part of them forms the nucleation core.

Introduction

The model of the atom proposed by Lord Kelvin initiated studies of knot topologies in physics as far ago as in 19th century (see Atiah, 1995 for review). The works of Tait, who was inspired by Kelvin's model and created the first table of knots, and Alexander, who developed an algebraic polynomial-based classification of knots, were first in these field. Formation of a knot may be conceived as winding of certain regions of the chain by its other regions through a random walk. Such model of a Brownian motion around an obstacle was considered in connection with a great number of physical phenomena, such as entangling of polymers (Edwards, 1967) and Solar magnetic lines (Berger, 1987); chiral diffusion in solids (Yakobson, Titov, 1988); and fractional statistics of particles (Belov et al., 1991). The pioneer works of Levy (1972) analyzing the random walk around an obstacle lead to a paradoxical result with a divergence of winding number. The underlying reason was a point dimension of the obstacle, when winding became eventual even as a result of infinitesimal steps. This shortcoming was later bypassed though computer simulation (Berger, 1987; Vologodskii, 1988) and by analytical solution of the problem (Yakobson, Titov, 1988). However, the interactions of parts of the molecule cannot be neglected in the case of biopolymers, and for this situation either the perturbation theory for weak interaction (Titov, 1997) or exact solution for a special type potential is known.

Until recently, a spontaneous knot formation during folding has been considered negligibly improbable, as reflected even in terminology. For example, minimization of the energy of non-knotted RNA and protein conformations is a standard problem, although a free chain without interactions would be highly "entangled" (Atiah, 1995). (In other words, biopolymers are implicitly believed topologically unrelaxed.) Another example is provided by evaluation of the quality of 16S RNA structure models, where the presence of knots was considered a negative criterion. Thus, the physical barriers for knot formation during folding of biopolymers for a long time were considered insurmountable (Schulz, Schirmer, 1979). This belief was exploded by a recent discovery of a protein with a pronounced knot (Taylor, 2000). The goal of this work was to design a cubic-lattice protein with a knotted conformation and study the specific pattern of its native contacts.

Methods

Geometry. A standard model of the protein whose chain is represented as a walk between vertices of a cubic lattice was selected for the study. As conventionally, only the conformations forming compact (cubic) structures were considered globular. Unlike the ring chain, it is impossible to determine strictly a knot in a chain with free ends. However, the chain ends may be virtually continued to enclose a ring. Here, a topological invariant—the number of crossings (taking the direction into account) of the chain with the surface of loop contour—characterizes the presence or absence of a knot. It is not difficult to see that a 27-mer protein (Fig. a) in this situation represents the minimal size chain capable of forming a knot (that is, displaying a nontrivial value of the topological invariant).

All the compact folding variants of this 27-mer were calculated. The conformation variants forming a knot were selected of them. One of all the knotted conformations (Fig.) was chosen for optimization.

Energy model. We used one of the most widespread interaction models, the so-called model of "perturbed homopolymer" (Shakhnovich, Gutin, 1990), that is, the model with the following Hamiltonian:

$$H=\frac{1}{2}\sum_i \xi_{ij}\delta_{ij} - \Delta,$$

where $\delta_{ij} = 1$ if the monomers *i* and *j* interact, that is, are coupled by an edge of the lattice and are not neighbors in the protein sequence, otherwise $\delta_{ij} = 0$; ξ_{ij} is the energy of interaction of the monomers *i* and *j*. Initially, ξ_{ij} were assumed random variables distributed normally. While designing, two probability moments were fixed: $\langle \xi_{ij} \rangle = 0$ and $\langle \xi_{ij}^2 \rangle = 1$ (Fig. b).

Algorithm

For the model described above, the sequence design consists in optimizing the matrix ξ_{ij} of contact energies. The design was performed in two steps: stability of the target structure was optimized at the first stage and its folding rate, at the second. At both stages, the sequences were evolved by genetic algorithm.

Thermodynamic optimization consisted in searching for such a set of ξ_{ij} that would provide a minimal energy of the target structure compared with all the alternative

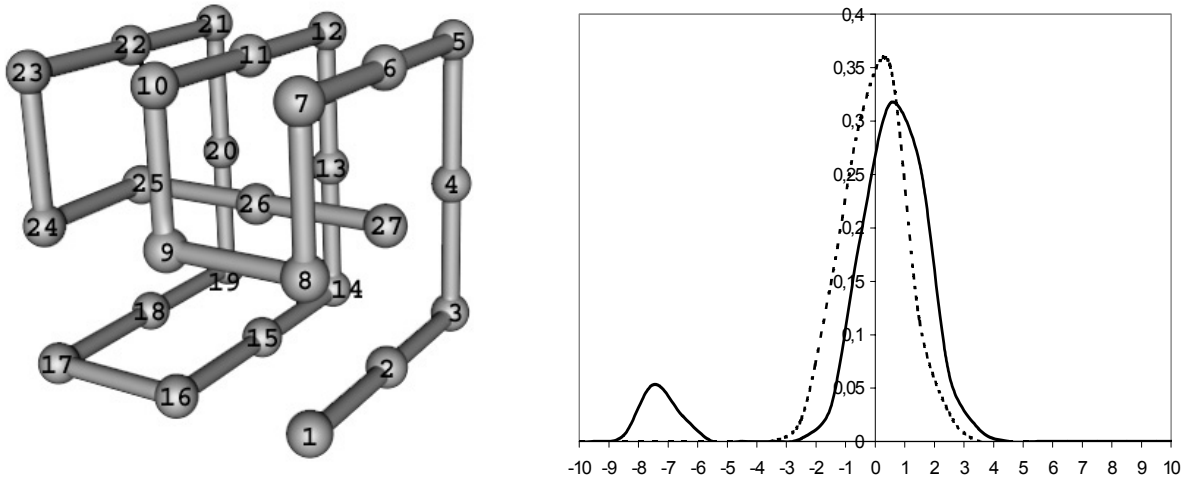


Fig. A compact conformation of 27-mer containing a knot: (**a**) the region 25–27 crosses the surface pulled over the loop 9–16 and (**b**) dynamics of evolution of the contact energies ξ_{ij} —distribution of matrix elements at the beginning (dotted line) and end of the (solid line) of the optimization.

compact folds. Practically, it appeared sufficient to minimize the energy of the target fold and then to make sure that it is superior compared with the alternative conformations for the given ξ_{ij} set. Several variants of the ξ_{ij} matrix were calculated. The other method for calculating ξ_{ij} matrix involved perturbing the initial ξ_{ij} matrix elements and controlling the energy gap between the target and alternative structures. Both methods gave similar results.

Kinetic optimization was the most laborious procedure and consisted in minimization of the rate of folding to the target structure. The folding kinetics was simulated using the Metropolis' algorithm (Metropolis et al., 1952) and two criteria of the folding rate: the number of steps required to form (a) only knot contacts (bonds between the monomer 26 and monomers 9, 11, 13, and 15) and (b) all the native contacts.

Results and Discussion

Interestingly, the number of knotted conformations of the 27-mer with a chain starting in a cube corner differs strongly from the number of knotted conformations coming from the midedge (recall that by virtue of the known theorem, there are no such compact folds of the 27-mer that start or end at the midface or center of the cube). As a result of thermodynamic optimization, all the native contacts appeared approximately similar in their value. They form a minor peak in Fig. b. It is natural, as these contacts contribute equally to the stability of the target fold and approximately equally to the stabilities of alterative folds. This situation resembles the Go's model of proteins (Ueda et al., 1977), which neglects the non-native interactions (the major peak of solid line in Fig. b).

In the kinetic optimization, the knot-forming contacts, that is, the contacts between monomer 26 and its neighbors and monomers 9 and 16, appeared most essential. Consequently, they were subjected to the first-turn (crude) optimization. However, they also contain most important contacts forming the nucleation core. The folding proceeds not via pulling the chain through loops, but through forming the contacts of monomer 26 with a part of the loop and subsequent locking of the loop with monomers 9 and 16. It would be next interesting to study whether this core is invariant while designing proteins homologous in their spatial structure. Or, in other words, to what degree the topology of the final structure determines the nucleation core and the folding route itself?

Acknowledgements

The work was supported in part by the Siberian Branch of the Russian Academy of Sciences (Integration Project № 65 and INTAS (grant № 2001-2126).

References

1. Ayiah M. (1995) Quantum physics and the topology of knots. Rev. Mod. Phys. 67 (4), 977-981.
2. Belov A.A. et al. (1991) The anion lattice gas. JETP. 100(7), 339-347.
3. Berger M. (1987) The random walk winding number problem: convergence to a diffusion process with excluded area. J. Phys. A.: Math. Gen. 20, 5949-5960.
4. Edwards S.F. (1967) Statistical mechanics with topological constraints. I. Proc. Phys. Soc. 91(573), 513-519.
5. Levy P. (1972) Stochastic processes. M.: Nauka.
6. Metropolis N., Rosenbluth A.W., Rosenbluth M.N., Teller A.N., Teller E., (1953) J. Chem. Phys. 21, 1087.
7. Schulz G. E., Schirmer R. H., (1979) Principles of Protein Structure. New York: Springer-Verlag.
8. Shakhnovich E., Gutin A. (1990) Implications of thermodynamics of protein folding for evolution of primary sequences. Nature. 346, 773-775.
9. Taylor W.R. (2000) Nature. 406, 916-919.
10. Titov I.I. (1997) PhD Thesis, Institute of Semiconductor Physics, Novosibirsk.
11. Ueda Y., Taketomi H., Go N., (1978) Biopolymers. 17, 1531.
12. Vologodskii A.V. (1988) Topology and Physical Properties of Circled DNA. M.: Nauka.
13. Yakobson B.I., Titov I.I. (1988) Kinetics of diffusion around screw dislocations. Latv. J. Phys. Tech. Sci. 1, 82-88.

DOMAIN STRUCTURE OF GLOBULAR PROTEINS AND DNA-PROTEIN INTERACTIONS

[1] Anashkina A.A., [2] Berezovsky I.N., [3] Namiot V.A., [4] Tumanyan V.G., [4*] Esipova N.G.

[1] Moscow Institute of Physics and Technology, Moscow, 141700, Russia
[2] Weizmann Institute of Science, Rehovot, 76100, Israel
[3] Research Institute of Nuclear Physics of Moscow State University, Moscow 117234, Russia
[4] Engelhardt Institute of Molecular Biology, RAS, Moscow, 119991, Russia
[*]The corresponding author. e-mail: nge@imb.imb.ac.ru

Key words: *DNA-protein recognition, protein domain structure, van der Waals interactions*

Resume

Motivation DNA-protein recognition is a challenge for experimentalists and theoreticians which try to explain these complex phenomena by applying experimental and theoretical methods. In order to shed light on specificity of the DNA-protein complexes it is necessary to investigate various types of interactions. Although hydrogen bonding and electrostatic interaction are often discussed in this respect there is some lack in analysis of van der Waals forces. Nevertheless, van der Waals interactions pattern is sufficient for deduction of domain structure of globular proteins. It is tempting to apply this kind of analysis to DNA-protein combining and estimate the relative role of van der Waals contacts in establishing of the proper structure.

Results By searching of energy domain structure on the basis of developed method there were studied interactions in twelve DNA-binding proteins and corresponding DNA-protein complexes. Characteristic van der Waals interactions (in particular DNA-protein interactions) were delineated and specific pattern for each type of complex was shown. It is demonstrated that resulting contribution of van der Waals interactions in energy of DNA-protein specific contacts is of same order of magnitude compared to interactions between domains in protein. Thus, reorganization of domain architectonics in the globule of a DNA-binding protein contributes mainly to stability and regulation of specific DNA-protein interactions.

Availability The program for van der Waals energy distribution estimation in DNA-protein contacts region for DNA-protein complexes is available on request from nastya@imb.ac.ru.

Introduction

DNA-binding proteins as a rule are built from distinct domains which may be readily traced by electron density distribution. DNA-protein interactions include Coulomb and van der Waals terms as well as hydrogen bonds. Since the number of contacts is limited, the energies of DNA-protein contacts will be similar to those for interdomain interactions. Hence the significant role of interdomain interactions and domain rearrangement of DNA-binding proteins in organization and controlling DNA-protein interface is obvious.

With the aim to illustrate this thesis it is necessary to study both stabilization of domain structure of DNA-binding globular proteins and role of attractive interactions in DNA-proteins interface.

In general, action of various processes on state and properties of biological structures for various levels of their organization is determined by energy balance between the energy generated in course of the processes and the minimal energy stabilizing elements of some level of the system. Let the condition $E_i >> E_{i+1}$ be satisfied, where E_i and E_{i+1} are characteristic energies of interactions between the elements of i-th and i+1-th levels of organization. Under this condition it is namely that level of biopolymer organization that undergoes alterations which is composed from elements with characteristic definite interaction energies. These interaction energies must be of same order of magnitude as the energy variations in the processes (in our case we mean the specific processes of DNA-protein interaction).

Van der Waals interactions must exist for any atom pair in condense medium. Van der Waals attractive forces originate from correlated fluctuations of atom dipole moments. Since the electric field from the dipole decreases as R^{-3}, as well as the electric field from the induced dipole, the interaction energy determined by the induced dipole at the place of the first dipole will be proportional to R^{-6}.

Since physical basis of the dispersional interactions is fluctuation of the dipole moments, characteristic fluctuation frequencies correspond to the frequencies of the UV radiation. Thus, in contrast to constant electrostatic field, which can be shielded by water with counter ions as well as charged groups of protein and DNA, van der Waals interactions are not shielded at all and occur in every pair of particles. In addition, these interactions are always attractive, and quite similar in magnitude because of the weak variation in local permittivity for protein or nucleic acid monomers.

Taking into account the above mentioned, one can estimate the complete dispersional energy as $E \sim \frac{N}{R^6}$, where N is the number of atoms pairs in the domain, R is typical size of the domain. Since $N \sim (nR^3)^2$, E does not decrease with increasing domain size, despite the increasing distances between interacting atoms in the pairs. As a result, $E \sim n^2$, where n is atom density in the domain. Thus, van der Waals energy can be expressed from the medium density n. In the regions of increased density, the van der Waals energy will be higher, and we can delineate a domain. See for explicit description (Berezovsky et al., 1999).

Table 1. Alterations in energy and cooperative properties of domains interacted with DNA.

Criterion	**Repressor protein**	**Criterion**	**Repressor protein with operator fragment**
	Trp-repressor (1jhg)		**(1tro)**
0.3Eo	**1:**8-41 **2:**43-48,50-108	**0.3-0.25Eo**	**1:**5-41 2:43-105
0.25-0.15Eo	**1:**8-41 2:43-48, 50-91 3:93-108	**0.2-015Eo**	**1:**5-41 2:43-48, 50-90, 92-105
	1:8-41 2:43-48,50-65,93-108		**1:**5-34, 36-41 2:43-48, 50-90
0.1Eo	**3:**67-76, 78-91	**0.1Eo**	02-105
	1:8-14,16-22, 2:24-34, 36-41		
	3:67-76,78-91 **4:**43-48, 50-		**1:**5-20 2:22-34,36-41
0.05Eo	65,93-97,99-108	**0.05Eo**	**3:**43-48,50-70,72-76,78-90
	Cro-protein (2cro)		**(3cro)**
0.3-0.25Eo	one domain protein	**0.3-0.25Eo**	one domain protein
0.20-0.15Eo	**1:**1-42, 44-64	**0.2-015Eo**	**1:**1-40 2:42-65
0.1Eo	**1:**1-14,44-64 2:16-42	**0.1Eo**	**1:**1-26,28-40 42-54,56-65
0.05Eo	**1:**1-14,16-25,27-32,34-42	**0.05Eo**	**1:**1-14,16-20,22-26, 28-32,34-40,42-48,
	2:44-48, 50-54		56-65
			2:50-54

Table 2. van der Waals interactions of DNA-protein complexes.

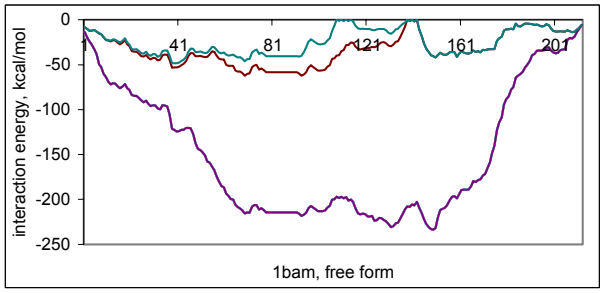
	Total van der Waals interactions for whole globula (kcal/mol)	Total van der Waals interactions with DNA (kcal/mol)
>1ESG: TYPE II RESTRICTION ENDONUCLEASE BAMHI BOUND **TO A NON-SPECIFIC DNA**	-2009.46	A-C: -9.247 A-D: -3.998
>1BAM: RESTRICTION ENDONUCLEASE BAMHI (E.C.3.1.21.4) **free form**	-1955.58	
>3BAM: RESTRICTION ENDONUCLEASE BAMHI COMPLEX WITH DNA AND MANGANESE IONS **(POST-REACTIVE COMPLEX)**	-1969.88	A-C: -32.194 A-D: -7.281 A-E: -45.141
>2BAM: RESTRICTION ENDONUCLEASE BAMHI COMPLEX WITH DNA AND CALCIUM IONS **(PRE-REACTIVE COMPLEX)**.	-2003.02	A-C: -38.528 A-D: -55.309
>1BHM:RESTRICTION **ENDONUCLEASE BAMHI COMPLEX WITH DNA**	-1855.52	A-D: -49.386 A-C: -25.517

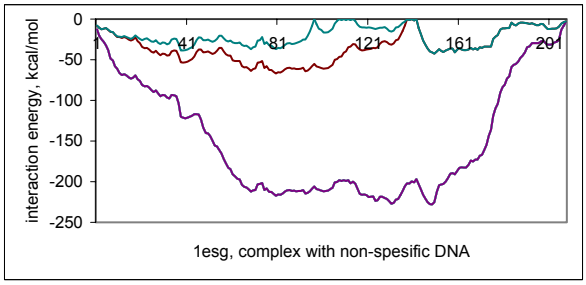
Determination of hierarchy of domain structures and distribution of van der Waals interactions energy in three-dimensional structure of DNA-binding proteins and their complexes was performed by the program whose algorithm is described in (Berezovsky et al., 1997). The method objectives are: a) to choose an approach for estimation of interactions energy in proteins and potential functions for calculation of nonbonded interaction energy; b) to compute distribution of nonbonded energy on some regions of protein molecule; c) to display the distribution of nonbonded interactions energy on various levels of protein structures and protein complexes structures hierarchy; d) to delineate energetically favourable regions composed of various super-secondary structure types.

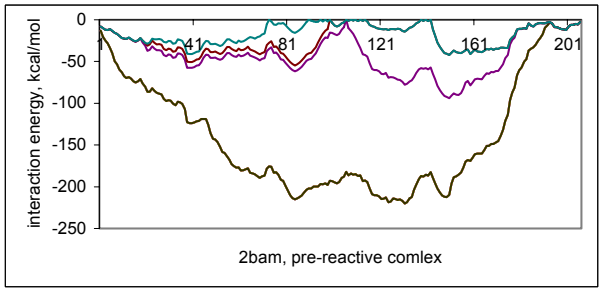
Results

With the aim to analyze the degree of protein domain structure influence on mechanisms of DNA-protein binding we study the hierarchy of interdomain interactions in DNA-binding proteins. These data are exhibited for *cro*-protein and *Trp*-repressor in Table 1. It can be seen that main alteration of pattern of van der Waals interactions energy in the protein globule

alone and in the DNA-protein complex is seen in interdomain contacts at various levels of energy hierarchy. Interestingly, in *Trp*-repressor-DNA complex the protein increases its level of it compactness beginning with the first hierarchy level. The overall number of domains rests unaffected despite of some variations in domains composition. In contrast, during *cro*-repressor-DNA complex formation the number of energy independent regions in protein globule varies simultaneously with alterations in interdomain interactions at various levels of energy hierarchy. Interrelations between contact energies in specific and non-specific complexes of endonucleases and DNA (Newman et al., 1994) were studied (see Table 2 and Figure 1-3). Amino acids involved in specific interaction are characterized by increased van der Waals energy (see Table 2). Van der Waals energy of specific DNA-protein contacts is close to interdomain energy of local protein regions. From Figure 1-3 one may conclude that the free protein and the protein in non-specific complex are similar. In specific complex the protein domain structure undergoes transformation.







Acknowledgments

The work was supported by the Russian Foundation for Basic Research (Grant 00-04-48351).

References

1. Berezovsky I.N., Tumanyan V.G., Esipova N.G. (1997) Representation of amino acid sequences in terms of interaction energy in protein globules. FEBS Lett. 418: 43-46.
2. Berezovsky I.N., Namiot V.N., Tumanyan V.G., Esipova N.G. (1999) Hierarchy of the interaction energy distribution in the spatial structure of globular proteins and the problem of domain definition. J. Biomol. Struct. Dyn. 17: 133-155.
3. Newman M., Strzelecka T., Dorner L.F., Schilkraut I., Aggarwal A.K. (1994) Structure of restriction endonuclease BamIII and its relationship to EcoRI. Nature. 368: 660-664.

CONTRIBUTION OF COORDINATED SUBSTITUTIONS TO THE CONSTANCY OF PHYSICOCHEMICAL PROPERTIES OF ATP-BINDING SITES IN PROTEIN KINASES

Afonnikov D.A.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: ada@bionet.nsc.ru

Key words: *amino acid sequences, coordinated substitutions, physicochemical properties of amino acids, protein kinases*

Summary

Motivation: It is known that the physicochemical properties of a protein that determine the specific folding of its polypeptide chain and functional features remain stable in the course of evolution. Search for such conserved features by analysis of homologous sequences may aid understanding the function, structure, and evolution of proteins under study. Coordinated substitutions of amino acid residues are one of the plausible mechanisms maintaining these features.

Results: This study deals with the contribution of coordinated amino acid substitutions to the constancy of several integral physicochemical properties of the ATP-binding loop in protein kinases. It has been shown that coordinated substitutions contribute to the conservation of some integral properties of the loop related to the charge, hydrophobicity, and frequency in the β-structure.

Introduction

Search for the most conserved features of proteins is one of the main tasks in the analysis of their homologous sequences. At the level of primary structure, this appears as conserved amino acid patterns. Moreover, proteins can be conservative with regard to their physicochemical properties depending on amino acid residues in several positions of the protein (integral characteristics). These characteristics are exemplified by the total volume of the hydrophobic core (Lim, Ptitsyn, 1970; Gerstein et al., 1994). The variance of an index is taken as a measure of its conservation.

One of the plausible mechanisms for maintaining integral characteristics is coordinated substitutions of amino acid residues. They are assumed to result from interaction between amino acid residues in a protein. The program package CRASP developed by Afonnikov et al. (2001) allows search for coordinated amino acid substitutions in families of homologous sequences and evaluation of their contribution to the conservation/variability of integral physicochemical characteristics of proteins.

In this study, the effect of coordinated amino acid substitutions on the constancy of some integral characteristics depending on the physicochemical properties of amino acids in the ATP-binding loop of protein kinases is analyzed. Twenty integral characteristics of amino acids have been investigated in this domain.

Methods and Algorithms

Integral physicochemical characteristics. An integral characteristic F_j, corresponding to the physicochemical property of amino acids j in a sequence of length L is considered to be the total value of this property in the positions of the sequence

$$F_j = \sum_{i=1}^{20} f_{ij} \ . \qquad (1)$$

The author analyzed 19 characteristics from the AAIndex database (Tomii, Kanehisa, 1996) and one characteristic not related to any properties: the ordinal numbers of amino acids. The names of amino acids and their codes in AAIndex are shown in Table 1. Weighing according to (Felsenstein, 1985) is used to take into account the evolutionary relations of sequences.

Criterion of the conservation of an integral characteristic. The contribution of amino acid substitutions to the conservation of integral characteristics is considered. For this purpose, the author used the criteria proposed in the earlier study (Afonnikov et al., 2001).

The variance of an integral characteristics $D(f_i)$ can be expressed as sum of two constituents: D_{var}, determined by the variability of protein positions, and D_{cov}, determined by coordinated amino acid substitutions:

$$D(F_j) = D_{var} + D_{cov}.$$

Note that D_{var} is always equal to or greater than zero, whereas D_{cov} may be positive, negative, or null. The last case can be used as the null hypothesis for verifying the significance of the contribution of coadaptive substitutions to $D(F_\alpha)$. In this case (for all $r_{ij} = 0$), the expected variance of the physicochemical characteristic $D_{exp}(F_\alpha)$ is equal to D_{var}.

$$D_{exp}(F_j) = D_{var} = \sum_{i=1}^{L} D(f_{ij}),$$

Coordinated substitutions contribute to the stability of the integral index F_j if

$$D(F_j) < D_{exp}(F_\alpha). \quad (2)$$

For validation of inequality (2), we may use a ratio of dispersions $\lambda = D_{exp}(F_j)/D(F_j)$, which, under the null hypothesis of equal dispersions, obeys the F distribution with $L(N-1)$ and $N-1$ degrees of freedom. Here, N is the number of sequences and L, number of positions (Selvin, 1998).

Additionally, the Monte Carlo technique is applied to estimate the statistical significance of the observed deviation of the parameter λ from unity. We generate normally distributed independent numbers with the mean and variance equal to their estimates at each position within the cluster and repeat this procedure N times. Then, we estimate the dispersion $D_{rand}(F_j)$ for such random samples. We repeat this procedure M=100,000 times, and count samples with $\lambda_{rand} > \lambda$, the ratio of this number to M, p_{MC}, is an estimate of $p(\lambda_{rand} > \lambda)$ under the null hypothesis. These values have also been compared to tabulated probabilities p_{tab}.

Implementation and Results

Sequence sample. The sample of protein kinase sequences from (Hanks, Quinn, 1991; http://www.sdsc.edu/kinases/pk_home.html) has been analyzed. Its size, after discharge of identical sequences, is 388 sequences. The ATP-binding loop is located at positions 43–64 of the protein cAPKa (Hanks, Quinn, 1991) and is formed by two β-strands (Fig.). Positions 43–62 are considered, because some of the sequences have deletions in the alignment positions corresponding to 63–64. Analyzed positions correspond to alignment positions 1–20. The number of degrees of freedom in the test (2) was $L(N-1)$=7740 and $N-1$=387.

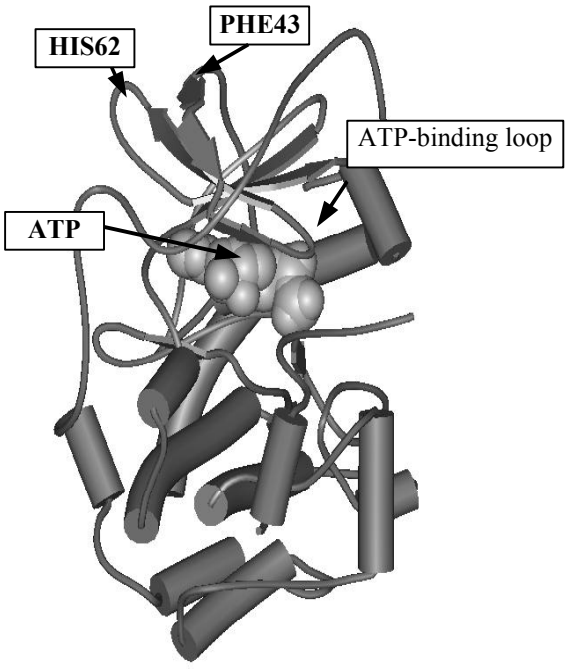


Fig. Schematic representation of the 3D structure of the catalytic domain of protein kinases. Arrows indicate the ATP-binding loop, loop N-terminal amino acid (Phe43), loop C-terminal amino acid (His62), and an ATP molecule.

The results of the analysis for 20 amino acid characteristics are shown in Table. They demonstrate that the contributions of coordinated substitutions to the conservation of different characteristics of the ATP-binding site differ considerably. The properties of amino acids are shown in the table in the decreasing order of the parameter p_{tab} A small value of this parameter shows that coordinated substitutions add to the conservation of integral characteristics. A greater value indicates that they increase variability.

The least values of p_{tab} and p_{MC} are characteristic of amino acid properties determining their hydrophobicity (Table, lines 1–6, 9), charge (isoelectric point, line 7), and occurrence in the β-structure (line 8). Hence, coordinated substitutions support the conservation of these amino acid characteristics in the ATP-binding loop.

On the contrary, coordinated substitutions do not contribute much to the variances of another group of characteristics (lines 10–19) at the error level of p=1%. Note that this group includes the index number of an amino acid, which has no physical sense (line 11). Analysis of the helix formation index (row 20) shows that coordinated substitutions increase the characteristic (1) variance significantly.

It is known that during evolution, amino acid sequences undergo selection pressure, which is directed to maintain the structure and function of a protein (Gerstein et al., 1994). For integral characteristics, selection pressure also depends on the functional and structural role of a particular characteristic. Therefore, concerning the ATP-binding loop, we may suggest that hydrophobic and electrostatic interactions are most important for this protein domain. They are likely to ensure binding of an ATP molecule. Moreover, the conservation of the index related to the frequency of an amino acid in the β-structure may reflect the role of the secondary structure of this domain in the β-conformation.

The role of other amino acid indices is less significant. This manifests itself in that they change the variance of F_j insignificantly or even increase it.

Table. Analysis of the contribution of coordinated substitutions to the variances of 20 integral characteristics calculated according to (1) for 20 physicochemical properties of amino acids. Names of the properties and their codes in the AAIndex database (Tomii, Kanehisa, 1996); variances of characteristics; their ratio λ to the expected value D_{exp} in the absence of coordinated substitutions; proportion p_{MC} of randomized samples in which $\lambda<\lambda_{rand}$; and evaluation of this value on the basis of the F distribution of p_{tab}

j	Physicochemical property	Code	D(F)	λ	$p_{MC}(\lambda<\lambda_{rand})$	$p_{tab}(\lambda<\lambda_{rand})$
1	Hydropathy (Kyte and Doolittle 1982).	KYTJ820101	159.59	1.53	0.00000	0.000000
2	HPLC parameter (Parker et al., 1986).	PARJ860101	567.2	1.51	0.00000	0.000000
3	Solvation free energy (Eisenberg-McLachlan, 1986).	EISD860101	26.13	1.47	0.00000	0.000000
4	Free energy of transfer to surface (Bull and Breese 1974).	BULH740101	12.50	1.45	0.00000	0.000001
5	Hydrophobicity (Eisenberg et al. 1984).	EISD840101	8.34	1.45	0.00000	0.000001
6	Hydrophilicity (Hopp and Woods 1981).	HOPT810101	65.03	1.40	0.00000	0.000008
7	Isoelectric point (Zimmerman et al., 1968).	ZIMJ680104	89.22	1.34	0.00003	0.000078
8	Normalized frequency of beta-sheet weights (Levitt, 1978).	LEVM780102	1.42	1.28	0.00033	0.000670
9	Average accessible surface area (Janin et al., 1978).	JANJ780101	14506.	1.25	0.00089	0,001822
10	Average flexibility indices (Bhaskaran and Ponnuswamy 1988).	BHAR880101	0.14	1.14	0.02868	0.042569
11	Amino acid index number .	-	1013.03	1.09	0.10091	0.128731
12	Normalized frequency for reverse turn with weights (Levitt 1978).	LEVM780103	2.75	1.07	0.17672	0.187593
13	Polarity (Zimmerman et al., 1968).	ZIMJ680103	12407.5	1.07	0.14008	0.187593
14	Free energy in beta-strand region (Munoz and Serrano et al., 1994).	MUNV940103	2.14	0.95	0.75417	0.765414
15	Refractivity (Jones, 1975).	MCMT64150101	1841.9	0.93	0.85991	0.878659
16	Normalized frequency of alpha-helix with weights (Levitt 1978).	LE16VM78011701	2.30	0.92	0.87097	0.878659
17	Free energy in alpha-helical region (Munoz and Serrano et al., 1994).	MUNV940102	1.38	0.86	0.98904	0.982859
18	Volume (Chothia, 1975).	CHOC750101	36522.9	0.86	0.98756	0.982859
19	Residue volume (Bigelow, 1967).	BIGC670101	11957.8	0.86	0.98403	0.982859
20	Helix formation parameters (delta G) (O'Neil and DeGrado, 1990).	ONEK900102	5.95	0.77	0.99997	0.999898

Reported in this work is analysis of several integral characteristics for the ATP-binding loop. The results show that coordinated amino acid substitutions contribute to the stability of the characteristics related to hydrophobicity, charge, and trend to formation of the β-structure of amino acids.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 01-07-90084); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); the Siberian Branch of the Russian Academy of Sciences (Integration Project № 65), US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Afonnikov D.A, Oshchepkov D.Yu., Kolchanov N.A. (2001). Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions. Bioinformatics. 17, 1035–1046.
2. Felsenstein J. (1985). Phylogenies and the comparative method. Amer. Nat. 125, 1–15.
3. Gerstein M., Sonnhammer E.L.L., Chothia C. (1994). Volume changes in protein evolution. J. Mol. Biol. 236, 1076–1078.
4. Hanks S., Quinn A.M. (1991). Protein kinase catalytic domain sequence database: Identification of conserved features of primary structure and classification of family members. Meth. Enzymol. 200, 38–62.
5. Lim V.I., Ptitsyn O.B. (1970). On the constancy of the hydrophobic nucleus volume in molecules of myoglobins and hemoglobins. Mol. Biol. (USSR). 4, 372–382.
6. Selvin S. (1998). F distribution. In: Armitage P., Colton T. (eds.). Encyclopedia of Biostatistics. Vol. 2. Chichester: John Wiley & Sons, 1469–1472.
7. Tomii K., Kanehisa M. (1996). Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 9, 27–36.

MUTATION RATE OF RIBOSOMAL PROTEINS AND THE 3D STRUCTURE OF THE SMALL RIBOSOMAL SUBUNIT

*[1] **Novichkov P.S.**, [1,2] **Gelfand M.S.**, [1,2] **Mironov A.A.**

[1] Integrated Genomics, P.O. Box 348, 117333, Moscow, Russia, e-mail: pnovichkov@integratedgenomics.ru
[2] GosNIIGenetika, 113545, Moscow, Russia
*Corresponding author

Key words: *mutation rate, ribosomal proteins*

Resume

Motivation: Ribosomal proteins are often used for construction of molecular phylogenetic trees. Thus it is interesting to analyze the conservation level of each ribosomal protein separately and to analyze the relation between conservation of a ribosomal protein and its functional importance.

Results: The relative mutation rates of the small subunit ribosomal proteins from gamma-proteobacteria were measured. It turned out that this group is not homogeneous, as the mutation rate of individual proteins can vary several-fold. The analysis of the spatial distribution of these proteins on the 3D structure of the small ribosomal subunit showed that it agrees with what is expected from functional considerations.

Introduction

Availability of a large number of complete genomes allows one to study the history of change of many proteins at once and to analyze the dependence between the mutation rate of a protein family and its functional importance.

Ribosomal proteins are one of the most conserved functional groups, and because of that they are often used for construction of molecular phylogenetic trees. However, this group is not homogeneous, as the mutation rate of individual proteins can vary several-fold.

Methods and Algorithms

We have measured the relative mutation rate of the small subunit ribosomal proteins from gamma-proteobacteria. The proteins were selected from the COG database (Tatusov et al., 2001). The relative mutation rate was derived from the similarity level using the Grishin formula (Grishin, 1995) and normalized by dividing by the arithmetic mean of the mutation rate of the fastest and the slowest proteins from this group (for details of the computations see Novichkov et al., 2002).

Implementation and Result

The result are shown in Table, where all proteins are divided into two groups, *slow proteins*, whose relative mutation rate is less than 1, and *fast proteins*, with the relative mutation rate exceeding 1. The fastest ones mutate almost at the same rate as an average (non-ribosomal) protein. We than mapped these proteins on the 3D structure of the small ribosomal subunit obtained from PDB entry 1FJG (Carter et al., 2000). The results are given in Fig.

The spatial distribution of the fast and slow proteins agrees with what is expected from functional considerations. Slow proteins are predominantly located in the "head" and "neck" of the small subunit, which are the most functionally important domains. Fast proteins are absolutely absent on the surface of contact of the small ribosomal subunit with the large subunit (Fig. a), moreover, the second most conserved protein S12 is located right in the center of this domain. Vice versa, fast proteins are all located on the convex back of the small ribosomal subunit (Fig. b). The two least conserved proteins S20 and S16, are located farthest from the functional center.

Thus in this case the mutation rate in a group of proteins forming one multi-protein complex nicely correlates with their functional and structural importance. Similar results were obtained in a preliminary study of the large ribosomal subunit and in other groups of bacteria.

This study was partially supported by RFBR, HHMI, INTAS, and LICR/CRDF.

Table. Relative mutation rate of proteins of the small ribosomal subunit.

COG name	COG ID	Relative rate
Slow proteins		
Ribosomal protein S10	COG0051	0.15
Ribosomal protein S12	COG0048	0.43
Ribosomal protein S11	COG0100	0.44
Ribosomal protein S18	COG0238	0.45
Ribosomal protein S19	COG0185	0.47
Ribosomal protein S5	COG0098	0.61
Ribosomal protein S3	COG0092	0.68
Ribosomal protein S4 and related proteins	COG0522	0.70
Ribosomal protein S7	COG0049	0.72
Ribosomal protein S9	COG0103	0.79
Ribosomal protein S21	COG0828	0.82
Ribosomal protein S14	COG0199	0.83
Ribosomal protein S8	COG0096	0.92
Ribosomal protein S13	COG0099	0.96
Ribosomal protein S17	COG0186	0.97
Fast proteins		
Ribosomal protein S2	COG0052	1.10
Ribosomal protein S1	COG0539	1.13
Ribosomal protein S6	COG0360	1.38
Ribosomal protein S15P/S13E	COG0184	1.50
Ribosomal protein S20	COG0268	1.81
Ribosomal protein S16	COG0228	1.85

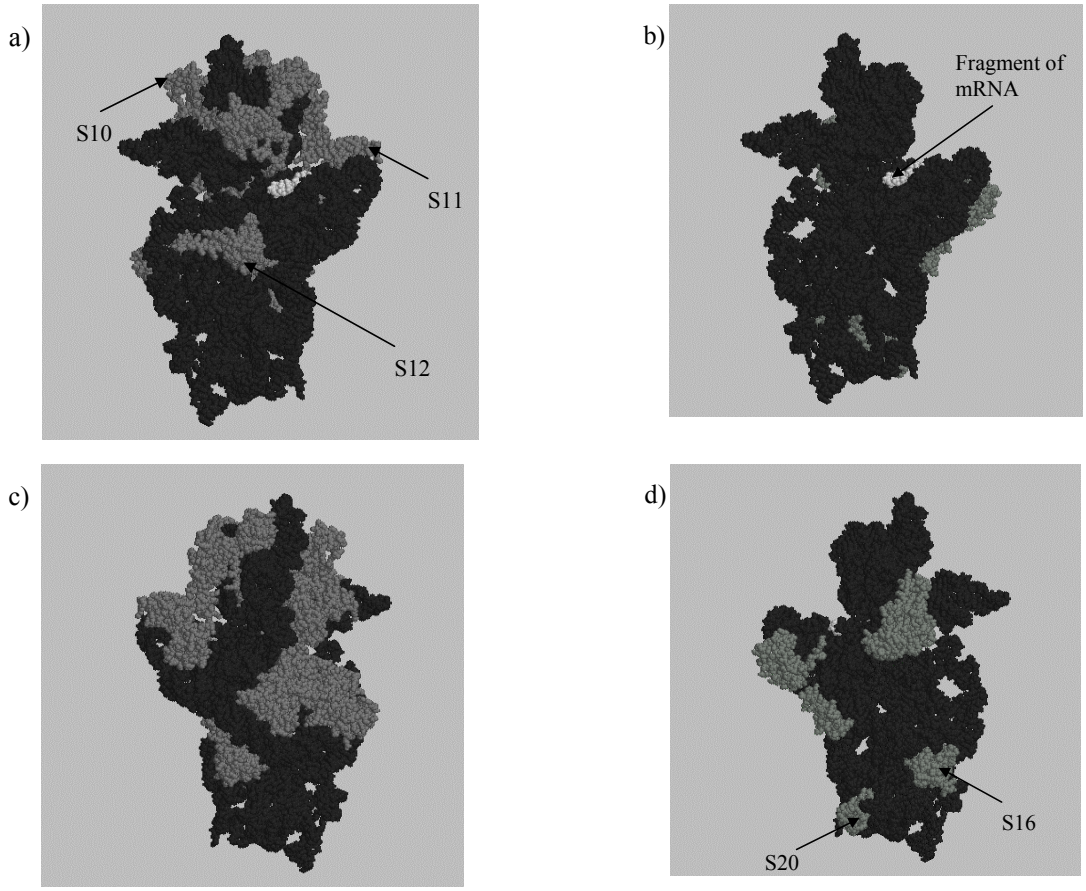


Fig. The small ribosomal subunit in two projections: frontal view (a and b) - surface of contact of the small subunit and the large subunit; back view (c and d) - the external side of the small ribosomal subunit. The 16S RNA is black, a fragment of mRNA is white, the ribosomal proteins are gray. Only slow proteins are shown in a and c; only fast proteins are shown in b and d.

References

1. Carter A.P., Clemons W.M., Brodersen D.E., Morgan-Warren R.J., Wimberly B.T., Ramakrishnan V. (2000) Functional insights from the structure of the 30S ribosomal subunit and its interactions with antibiotics. Nature. 407, 340.
2. Grishin N.V. (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J. Mol. Evol. 41, 675-679.
3. Novichkov P.S., Gelfand M.S., Mironov A.A. (2002) Relative mutation rate of bacterial proteins and prediction of the distance between orthologous genes // Proc. of the 3nd Int. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk.
4. Tatusov R.L., Natale D.A., Garkavtsev I.V., Tatusova T.A., Shankavaram U.T., Rao B.S., Kiryutin B., Galperin M.Y., Fedorova N.D., Koonin E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl. Acids Res. 29, 22-28.

RELATIVE MUTATION RATE OF BACTERIAL PROTEINS AND PREDICTION OF THE DISTANCE BETWEEN ORTHOLOGOUS GENES

*[1] **Novichkov P.S.**, [1,2] **Gelfand M.S.**, [1,2] **Mironov A.A.**

[1] Integrated Genomics, P.O. Box 348, Moscow, 117333, e-mail: pnovichkov@integratedgenomics.ru
[2] GosNIIGenetika, Moscow, 113545
*Corresponding author

Key words: *mutation rate, orthologs*

Resume

Motivation: The methods of comparative genomics largely rely on analysis of orthologous genes. Thus, identification of orthologs in completely sequenced genomes is an important problem. In particular, it is necessary to distinguish orthologs from paralogous genes, whose divergence pre-dates the specitation event, and from recent horizontally transferred genes.

Results: We have tested the hypothesis that the distance between orthologous genes can be decomposed as the product of two terms, one of which depends only on the protein family, and the other depends only on the pair of analyzed genomes.

*Availability***:** http://212.48.144.189/rmr/index.jsp

Introduction

Identification of orthologs is one of the basic ingredients of genomic analysis. In many cases this is done by variants of the "bidirectional best hit" (BET) criterion (Tatusov et al., 2001): orthologs are defined as pairs of genes that are closest relative of each other in two considered genomes.

However, this criterion is not sufficient. Firstly, genes may form a BET even if they arose from duplication preceding the specitation event: it happens if each lineage lost one of the copies in the descendant genomes, and the lost genes belonged to different branches. Secondly, if a large multigene family is considered, in which the orthology relationships cannot be resolved even by careful analysis of phylogenetic trees, still it is simple to show that at least one pair of genes will form a BET according to purely formal reasons. Finally, genes may be close due to horizontal transfer from one genome to the other, or to both genomes from a common sources.

In the former two cases the similarity between the genes forming the BET will be lower than expected for this pair of genomes, whereas in the last case it will be higher than expected. However, although intuitively it is clear what is the expected similarity between orthologous genes from any given genome pair, it is also obvious that this value depends not only on the genomes, but also on the analyzed proteins. Here we test the hypothesis that the similarity (more exactly, the distance) between two representatives of a protein P_i in two genomes A and B, that is, the distance between orthologs A_i and B_i (from A and B respectively), can be expressed as the product of two terms, the first of which depending only on the protein family, and the second one depending only on the pair of genomes: $d(A_i,B_i) = V_i*D(A,B)$. Here V_i is the mutation rate of the protein family P_i, D is the distance between the genomes A and B. If correct, this means that the mutation rate of a protein does not depend on the genome, and that it is sufficient to measure once the distance between the genomes and the mutation rates for all protein families, in order to have the expected distance for any BET, if this BET indeed represents a pair of orthologs.

Here we systematically test this conjecture. As we do not have any independent estimate of the divergence times for bacterial genomes, we use the relative mutation rate (RMR), selecting one protein family as the basic one and comparing the distance between all orthologs to the distance between representatives of the basic family in the same genomes.

Methods and Algorithms

Consider two genomes, A and B. Chose a protein P_0 as the basic one. The relative mutation rate (RMR) of an arbitrary protein P_i relative to P_0 is defined as the ratio between distances of representatives of P_i and P_0 in these genomes (resp. A_i, B_i, A_0, B_0) measured in the number of mutations per position:

$$V_i = \frac{d(A_i, B_i)}{d(A_0, B_0)} \quad (1)$$

Given a large number of genomes containing representatives of P_i and P_0, one can measure RMR at diverse time intervals and thus to obtain statistically valid estimates.

The data were taken from the COG database (Tatusov et al., 2001). Two groups of genomes were considered, gamma-proteobacteria and Bacillus/Clostridium group (Table).

The distance between orthologs d was estimated given the computed number of identical amino acids in the alignment q using the Grishin formula (Grishin, 1995)

$$q = \frac{1 - e^{-2d}}{2d} \tag{2}$$

Table. Groups of genomes.

Gamma-proteobacteria		Bacilus/Clostridium group	
Genome identifier	Genome	Genome identifier	Genome
EcZ	Escherichia coli O157	Bha	Bacillus halodurans
Eco	Escherichia coli K12	Bsu	Bacillus subtilis
Hin	Haemophilus influenzae	Lla	Lactococcus lactis
Pmu	Pasteurella multocida	Spy	Streptococcus pyogenes
Vch	Vibrio cholerae	xLTm	Listeria monocytogenes
xSLtm	Salmonella typhimurium	xLTi	Listeria innocua
xSLt	Salmonella typhi	xSpn	Streptococus pneumoniae
xYP	Yersinia pestis	xSTam	Staphilococcus aureus Mu50
		xSTan	Staphilococcus aureus N315
		xCLa	Clostridium acetobutylicum

Note. Clusters of orthologous groups were extended by completely sequenced genomes missing in the COG database. Identifiers of these genomes start with "x"

Assume that protein P and the basic protein have constant mutation rates v and v_0, respectively. Then for distances between orthologous proteins, one can write a system of equations

$$\begin{cases} d_k = v * t_k + \varepsilon_k \\ d_{0k} = v_0 * t_k + \delta_k \end{cases} \tag{3}$$

where d_k and d_{0k} are the distances between orthologs representing the protein P and the basic protein; ε_k and δ_k are the errors. Applying the regression analysis minimizing

$$\frac{\varepsilon_k^2 + \delta_k^2}{\sqrt{d_k^2 + d_{0k}^2}} \tag{4}$$

one gets the estimate of the relative rate $V=v/v_0$. The estimate quality is measured by the correlation coefficient

$$CC = \frac{\sum d_k d_{0k}}{\sqrt{\sum d_k^2 \sum d_{0k}^2}} \tag{5}$$

As the basic protein, we used COG0013, Alanyl-tRNA synthetase that is present in all genomes in the COG database and does not contain paralogs. Its mutation rate is roughly equal to the average mutation rate of all proteins (data not shown).

Implementation and Results

Analysis of the RMR in two considered groups of genomes shows that in most cases the experimental data can be well approximated by a straight line coming through the zero point (0,0). The tangent of this line is the estimate of the RMR for the given protein family. Fig. 1 presents the distribution of the correlation coefficient in the considered groups of genomes. One can see that the correlation coefficient exceeds 0.97 in approximately 80% of proteins from gamma-proteobacteria and about 70% of proteins in the Bacillus/Clostridium group.

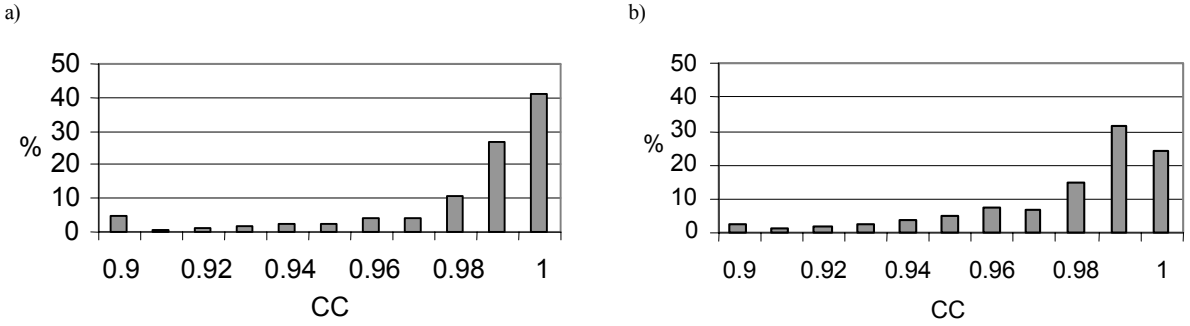


Fig. 1. Histogram of correlation coefficient CC. (a) gamma-proteobacteria; (b) Bacillus/Clostridium group. In both cases, only complete COGs were considered, that is, COGs having at least one representative in each genome from the given group. Vertical axis: % of proteins.

Thus, for majority of proteins the RMR defined by formula (1) indeed is constant for the given group of genomes, that is, depends only on the protein. Defining the distance between two genomes A and B through the distance of the representatives of the basic protein in these genomes: $D(A,B)=d(A_0,B_0)$, one finally arrive at

$$d(A_i,B_i) = V_i * d(A_0,B_0).$$

where the first term depends only on the protein family, and the second term depends only on the pair of genomes.

Analysis of RMR in gamma-proteobacteria shows rather wide distribution (Fig. 2). The slowest proteins, as expected, were ribosomal proteins and proteins involved in translation. The most conserved protein was a ribosomal protein S10 with RMR=0.07. The second most conserved protein is the translation initiation factor IF-1, with RMR=0.17. Among the most conserved proteins are membrane and secreted proteins with RMR reaching 4 and even more. Thus, genomes of gamma-proteobacteria contain proteins with RMR differing more than 20-fold.

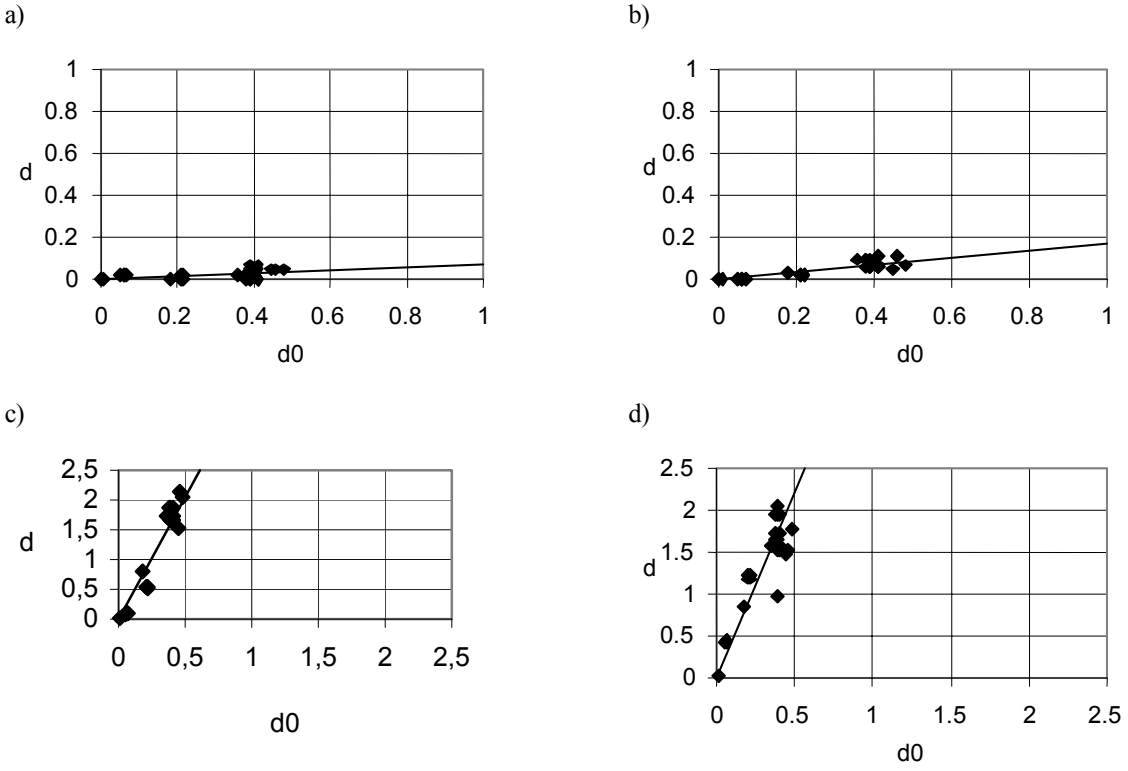


Fig. 2. Scatterplots of distances between orthologous proteins. Each dot corresponds to a pair of genomes. Horizontal axis: the distance for the basic protein. Vertical axis: the distance for the current protein: (a) ribosomal protein S10 (COG0051); (b) translation initiation factor IF-1 (COG0361); (c) general pathways protein F (COG1459); (d) uncharacterized lipoprotein (COG3317).

The range of RMR in the Bacillus/Clostridium group has also a wide spectrum, although it is narrower, mostly due to the absence of COGs with extremally large RMR. It is likely due to the absence of the external membrane in bacteria of this group, as the proteins of the external membrane and periplasm are the least conserved.

References

1. Grishin N.V. (1995) Estimation of the number of amino acid substitutions per site when the substitution rate varies among sites. J. Mol. Evol. 41, 675-679.
2. Tatusov R.L., Natale D.A., Garkavtsev I.V., Tatusova T.A., Shankavaram U.T., Rao B.S., Kiryutin B., Galperin M.Y., Fedorova N.D., Koonin E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucl. Acids Res. 29, 22-8.

STUDY OF CD150 CYTOPLASMIC TAIL INTERACTIONS WITH SH2-DOMAINS

Akimov Y.M., * Sidorenko S.P.

Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology, Kyiv, Ukraine, e-mail: svetasid@onconet.kiev.ua
*Corresponding author

Key words: ***protein structure, threading, docking, signal transduction pathways***

Resume

Motivation: A divergent functions of cell surface receptor CD150 depend on unique structure of its cytoplasmic tail (CD150ct) which could bind SH2-containing molecules. The modeling of CD150 interactions with different SH2-containing molecules will help to understand how CD150 initiates different signal transduction pathways.

Results: A 3D-model of CD150ct was made. This model was used for docking with SH2-containing proteins SH2D1A and SHP-2. We found preferential binding of SH2D1A to pY281, and SHP-2 to pY327 in CD150ct. Also we showed that binding of SH2D1A to Y281 in CD150ct changed the conformation of CD150ct and expose Y307 to phosphorylation. Docking of the SH2-containing molecules to CD150 demonstrated possible combinations of binding with the molecules that links this receptor to signal transduction pathways.

Introduction

Cell surface receptor CD150 (IPO-3/SLAM) belongs to CD2 family receptors of Ig superfamily (Sidorenko S.P. and Clark E.A., 1993; Cocks B.G. et al., 1995). Extracellular domain of CD150 includes 209 residues and forms two Ig-like domains. CD150 cytoplasmic tail (CD150ct) contains 77 residues. There are 4 tyrosine residues in CD150ct (269, 281, 307 and 327) two of which are within immunoreceptor tyrosine-based switch motifs (ITSM): TIYAQV (Y281-motif) and TVYASV (Y327-motif). SH2-containing tyrosine phosphatase 2 (SHP-2) and SH2-containing adaptor protein SH2D1A (both are cytoplasmic molecules) coprecipitate with CD150, and use the same binding sites within CD150ct: Y281 and Y327 (Poy et al., 1999; Shlapatska et al., 2001). SH2D1A has been intensively studied due to its involvement to X-linked lymphoproliferative disease (XLP), a rare inherited immunodeficiency, characterized by benign or malignant proliferation of lymphocytes, histiocytosis and alterations in serum immunoglobulin concentrations. Gene encoded this small adaptor protein is altered or deleted in XLP (Sayos et al., 1998). A number of missense mutations within the limits of the SH2 domain of SH2D1A directly implicate it in the pathogenesis of XLP. Despite of the known crystal structure of SH2D1A and SHP-2, mechanism of interactions between CD150, SH2D1A and SHP-2 is still unclear. Also, the spatial structure of CD150ct is not yet determined experimentally. The general aim of this work was to clarify interactions of SH2-domains of SH2D1A and SHP-2 with CD150ct on structural level. The specific tasks were: 1) to build 3D-model of CD150ct; 2) to perform CD150ct-SH2D1A docking; 3) to perform CD150ct-SHP-2 docking.

Methods

PDB does not contain protein 3D-structures that could be used as templates for CD150ct homology modeling. That is why CD150ct model was built using threading approach on FUGUE (http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html) and SAUSAGE (http://www.embl-heidelberg.de/predictprotein/submit_meta.html) servers on 2gn5-template of bacteriophage fd DNA-binding protein. Substitution of matrix residues, phosphorylation and energy optimization were made in HyperChem 6.0 package (http://www.hyper.com/products/default.htm). Studies of CD150ct-SH2-domains interactions were obtained by macromolecular docking method using Hex 2.4 program (http://www.biochem.abdn.ac.uk/hex/). Estimation criterions for generated complexes were the experimental data of our laboratory, and X-ray structures of SH2-domains in the complex with different phosphorylated and nonphosphorylated peptides including CD150ct fragment with Y281-motif: PDB entries 1d4t, 1d4w (Poy et al., 1999), 1aya (Lee et al., 1994). For predicting protein interface interactions we used Protein-Protein Interaction Server (http://www.biochem.ucl.ac.uk/bsm/PP/server/index.html). For phosphorylation sites prediction we used NetPhos 2.0 server (http://www.cbs.dtu.dk/services/NetPhos/).

Results and Discussion

CD150ct model after optimization remains in 2gn5 V-like conformation (RMS between model and template –1,76Å) – it has mainly beta OB-fold with open barrel architecture. In our model both ITSM-motifs are located each in one V-form "branch" (Fig. 1). Two halves of CD150ct are 43,3Å (Y281–containing branch) and 34,5Å (Y327-branch) length. Three tyrosine residues (281, 307 and 327) have great phosphorylation potential, as well as threonine 271 and 305 and serine 317. We consider our V-like model illustrates CD150ct bound state.

In this work we docked V-form of CD150ct with "open" active forms of SH2D1A adaptor protein and tyrosine-phosphatase SHP-2. Main difference between "open" and "closed" conformations in SH2D1A and SHP-2 SH2-domains is distance between BG and EF loops. In closed form it is 7,3Å and in open form it varies from 11,3Å to 12,3 Å. Open SH2D1A form was taken from 1d4t structure. SHP-2 was crystallized in inactive conformation with "closed" N-terminal SH2-domain, so for docking SHP-2 with V-form we transplanted an open N-terminal SH2-domain (1aya) into SHP-2 structure.

Performing CD150ct-SH2D1A docking we got complex with Y281-motif orientation, structurally identical to CD150ct-peptide orientation in 1d4t (RMS = 0,3Å). We found that Y281-motif docks to SH2D1A peptide-binding cleft on SH2-domain, formed mainly by βD, βC strands and αB spiral. SH2D1A binds both tyrosine phosphorylated and non-phosphorylated Y281-motif (Y327-motif binds only in phosphorylated form). Obligatory condition for CD150ct-SH2-domain interaction is BG and EF loops participation in peptide (and V-form) binding. Y281 is dipped into cavity on SH2-domain, formed by βC strand, αA spiral, and BC and AA loops. Such complex is formed even in pY327-CD150ct-SH2D1A docking (Fig. 1). It has interface surface area 806.67 $Å^2$ (~15% of total surface area). Interface gap volume is 4677.28 $Å^3$. Complex interface is stabilized by five H-bonds.

Mechanisms of Y281 and Y327-motifs binding to SH2D1A are different. pY327 docks to phosphopeptide binding cleft on SH2-domain (βC and βD strands) and interacts with BG-EF loops (Fig. 1). Y327-motif has greater affinity to SH2 domain in phosphorylated form: it has larger interface area (1169.34 $Å^2$), lower gap volume (3487.54 $Å^3$), it also stabilized by five H-bonds plus two salt bridges. But if Y281 is phosphorylated, SH2-domain of SH2D1A would have higher affinity to pY281 then to pY381. When Y281 is bound to SH2D1A, second SH2-domain weakly binds pY327 due to competition with H-bonds between pY327 in CD150ct, D48 and Y52 (weaker) in SH2D1A docked to Y281. SH2D1A can bind Y327-motif both in free (unbound) CD150ct and within CD150ctY281-SH2D1A complex, raising a possibility of CD150ct interaction with two SH2D1A molecules simultaneously. However, experimental data showed that complex of CD150 with SH2D1A also binds SH2-containing inositol phosphatase SHIP, which requires pY327 for binding.

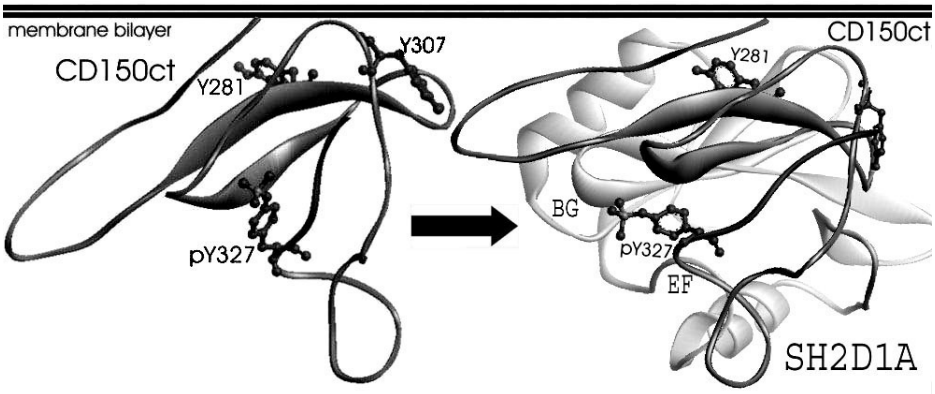


Fig. 1. CD150ct free (left) and SH2D1A binding to CD150ct (right). SH2D1A docked to Y281-motif (non-phosphorylated). pY327 is partially interacting with SH2-domain.

It was shown that CD150ct-SH2D1A interaction facilitates tyrosine phosphorylation of CD150ct (Latour et al., 2001). In our model Y307 located in V-form junction, and its location corresponds to the beginning of cytoplasmic domain, which enters the membrane. It means that probability of Y307 phosphorylation in CD150ct free (unbound) state is very low. SH2-domains can't interact properly with such a "corner-to-membrane" position of Y307 in CD150ct. Flexibility of first few residues in CD150ct allows SH2D1A after binding to Y281 to change a loop conformation and to turn Y307 into "corner-out-of-membrane" position (Fig. 1, right). This makes Y307 available for phosphorylation and then for the SH2-binding.

Tyrosine phosphatase SHP-2 has two SH2-domains: N- and C-terminal. After aligning SH2D1A, N- and C-terminal SH2-domains sequences it was revealed that SH2D1A is much more closely related to N-SH2 domain of SHP-2 (26% identical and 40% similar residues) than to C-SH2 (7% of identical and 16% of similar residues). After docking SHP-2 with pY281 and/or pY327 in CD150ct, we got complexes, in which pY327 binds N-terminal SH2-domain the way similar to SH2D1A (Fig. 2). Interface area in this complex is 938,33 $Å^2$, it includes 43,88% of all polar atoms. Complex interface is stabilized by three H-bonds. Docking showed that even in CD150ct with both Y phosphorylated (pY281 and pY327) SHP-2 more readily binds pY327. Probably, SHP-2 has higher affinity to pY327 than to pY281. This is supported by our experimental data. On the contrary, SH2D1A has higher affinity to pY281 than to pY327.

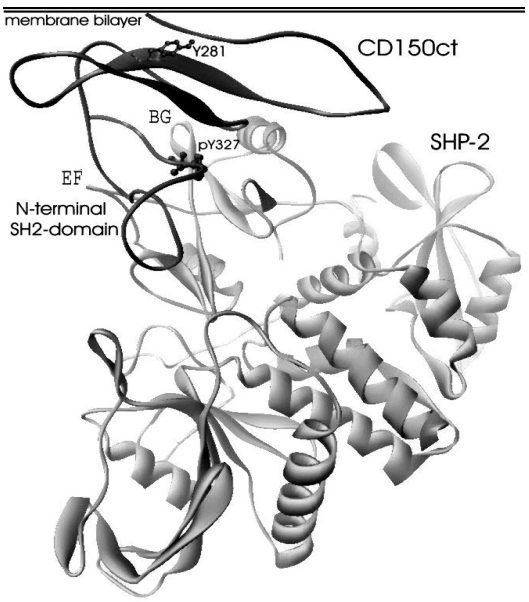


Fig. 2. SHP-2 bound to CD150ct. N-terminal SH2-domain docked to pY327.

Taken together, using Hex 2.4, we have shown different possibilities for SH2-domains docking to CD150ct: 1) Simultaneous binding of two molecules of SH2D1A to CD150ct theoretically is possible but is not supported by experimental data. 2) SH2D1A and SHP-2 could dock CD150 simultaneously, with preferential binding of SH2D1A to Y281, and SHP-2 to Y327 in CD150ct. 3). In the absence of SH2D1A, CD150ct potentially could bind two SHP-2 molecules.

Acknowlegements

The work was supported by CRDF grant UB2-531.

References

1. Cocks B.G., Chang C.C., Carballido J.M., Yssel H., de Vries J.E., Aversa G. (1995) A novel receptor involved in T-cell activation. Nature. 376, 260-263.
2. Hof P., Pluskey S., Dhe-Paganon S., Eck M. J., Shoelson S.E. (1998) Crystal structure of the tyrosine phosphatase SHP-2. Cell. 92, 441-453.
3. Lee C.H., Kominos D., Jacques S., Margolis B., Schlessinger J., Shoelson S.E., Kuriyan J. (1994) Crystal structures of peptide complexes of the amino-terminal SH2 domain of the Syp tyrosine phosphatase. Structure. 2, 423-438.
4. Poy F., Yaffe M.B., Sayos J., Saxena K., Morra M., Sumegi J., Cantley L.C., Terhorst C., Eck M.J. (1999). Crystal structures of the XLP protein SAP reveal a class of SH2 domains with extended, phosphotyrosine-independent sequence recognition. Mol. Cell. 4, 555-561.
5. Sayos J., Wu C., Morra M., Wang N., Zhang X., Allen D., De Vries S.E., Aversa G., Terhorst C. (1998) The X-linked lymphoproliferative-disease gene product SAP regulates signals induced through the co-receptor SLAM. Nature. 395,462-469.
6. Shlapatska L.M., Mikhalap S.V., Berdova A.G., Zelensky O.M., Yun T.J., Nichols K.E., Clark E.A., Sidorenko S.P. (2001) CD150 association with either the SH2-containing inositol phosphatase or the SH2-containing protein tyrosine phosphatase is regulated by the adaptor protein SH2D1A. J. Immunol. 166, 5480-5487.
7. Sidorenko S.P., Clark E.A. (1993) Characterization of cell surface glycoprotein IPO-3, expressed on activated human B and T lymphocytes. J. Immunol. 151, 4614-4624.

PROTEIN FAMILY PATTERNS BANK PROF_PAT IS WORTHWHILE RIVAL TO WORLD-KNOWN "SECONDARY" BANKS

**** Nizolenko L.Ph., Bachinsky A.G., Yarigin A.A., Naumochkin A.N.***

SRC "Vector", Koltsovo, Novosibirsk, Russia, e-mail: lilnizolenko@mail.ru
*Corresponding author

Key words: *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 appeared recent years. Purpose of current study was to demonstrate it to be not only as good as known analogs but surpasses them on some features.

Results: Using 20 amino acid sequences from TrEMBL bank with no description, Prof_Pat demonstrated specificity as good as the best world-known "secondary" banks do. At the same time, its completeness and variety of included proteins are significantly more, and its search speed is 3-10 times higher then those of any foreign protein family bank we used for comparison.

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

A number of "secondary data banks" of sites (patterns, blocks, motifs) in the groups of related proteins, which are representative of a protein family as a whole are widely used recent years. They are used both to identify new proteins and to refine structural and functional properties of those already known. How can investigator choose the most appropriate one? In the most cases (excluding PRINTS), the databases are universal and can be used for very different purposes. We have compared world-known banks completeness, specificity and search speed with one another and with our bank Prof_Pat, to help in this task.

Methods and Algorithms

Protein family patterns, the bank of this patterns PROF_PAT and flexible fast search program were created using original technology (Bachinsky et al., 1997, 2000). The version of Prof_Pat 1.8 constructed on the basis of the 40th release of the SWISS-PROT bank and 18th release of TREMBL, contains patterns of 33718 groups of related proteins including more then 217000 amino acid sequences.

20 amino acid sequences were selected from Trembl's file "cumulative_dat" of 17.12.2001. The sequences belong to different species from Homo sapience to bacteria and viruses. The only criterion of selection was the absence in the field "DE" any description excepting short name of open reading frame.

The parameters of comparison of sequences with all databanks, including Prof_Pat, were standard, offered by authors in their web-sites (see Table 1).

Implementation

We have compared 9 world-known banks with one another and with our bank Prof_Pat for completeness, specificity and search speed using 20 amino acid sequences from Trembl's file "cumulative_dat" of 17.12.2001, whose function and (or) relationships still are not described. The sequences were examined on-line using 10 "secondary" protein banks with standard parameters, offered by databank's authors in their web-sites. Prof_pat was created for distance relation discovery, so the sequences were compared with Prof_Pat bank with similarity level 70% too. Results of this examination together with some other features of databanks presented in the Table 1.

It is obvious that search speed of Prof_Pat 3-10 times higher than search speed of any other protein family bank, mainly, because it is able to examine large groups of protein sequences rather than just few of them. Direct comparison of this set of 20 sequences using Prof_Pat without Internet takes less then 2 minutes. Thus data input and transfer of results took about half of time required, and that feature of Prof_Pat becomes very important. Bank Interpro can work with a set of sequences, however, spends essentially more time for these comparisons (Table 1).

As to specificity, the Table 1 shows all banks to be similar. Prosite and Prints are the exception. Prosite recognizes all 20 sequences, but it analyses sometimes very short fragments, that say nothing about protein function or relationships. Prints was created for special tasks and therefore recognizes only 3 sequences.

Hypothetical proteins, which function is uniformly defined by 9 databanks (except Prints), are presented in the Table 2.

Table 1. Comparison of completeness, specificity and search speed of world-known banks.

The name of the bank	Release, data	Number of patterns (entries, families)	Number of motifs	The time of search (min)	The number of positive results	The number of accurate results[1]	The source of data
Prof_Pat	1.8. 12.2001	33718	515778	3-4	14(15[2])	14(15[2])	http://wwwmgs.bionet.nsc.ru/mgs/programs/prof_pat/
Prosite	16.51 11.2001	1104	1494	11	20	?	http://www.expasy.ch/prosite/
PFAM	6.6 08.2001	3071	-	20	14	12	http://www.sanger.ac.uk/Pfam/
Prints	32.0 09.2001	1600	9800	8-9	3	3	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
Interpro	4.0 11.2001	4691	-	20-25	12	11	http://www.ebi.ac.uk/interpro/
Blocks	13.0 08.2001	2101	8656	24-25	20(12[3])	14(6[3])	http://blocks.fhcrc.org/blocks/
Sbase	9.0 12.2001	3164	-	Email only	17	8	http://www.icgeb.trieste.it/sbase/
Tigrfams	1.2 08.2001	1109	-	40	11	10	http://www.tigr.org/TIGRFAMs/
Emotifs	2001	15893	70297	9-10	7	7	http://motif.stanford.edu/identify
IproClass	1.1 12.2001	34,780 PIR superfam. 100,000 families	-	20-25	16(13[4])	14(11[4])	http://pir.georgetown.edu/pirwww/dbinfo/dbinfo.html

1 - Sequences, more then 20% similar to their presumptive Swiss-Prot/Trembl relatives; 2 - Comparison with similarity level 70%; 3 - Sequences identified by one half (or more) blocks of the family; 4 - 3 sequences were identified by "one-member-family".

Table 2. Hypothetical proteins, which function is uniformly defined by 9 databanks (except Prints).

Sequence	Description	Presumptive function
AAC24084	12M4.9 PROTEIN Arabidopsis thaliana	Histone H3*
AAL09790	AT3G07110/T1B9_24 Arabidopsis thaliana	Ribosomal protein L13
AAH12642	PROTEIN FOR MGC:13744 Mus musculus	Glutaredoxin*
AAK93536	SD05956P Drosophila melanogaster	RNA 3'-terminal phosphate cyclase
AAK93457	LP01967P Drosophila melanogaster	Queuine tRNA-ribosyltransferase
CAA09651	GRA-ORF6 PROTEIN Streptomyces violaceorube	Short chain dehydrogenase*
AAB18084	ORF_O136 Escherichia coli	Transposase
AAA48264	ORF8CDS Vaccinia virus	Carbonic anhydrase

* Hypothetical proteins, which function was defined by Prints too.

In some cases different banks ascribe different function to the same sequences. We have compared the sequences with proteins of the 40th release of Swiss-Prot and 18th release of Trembl directly and found them to be essentially more similar to proteins from Prof_Pat families found, then to proteins, described as their relatives by other banks (Fig.).

In the 6 cases amino acid sequences do not show any similarity with patterns of the bank PROF_PAT, but are recognized by some another bank. But all of them demonstrate no more then 17% similarity level when compared with their presumptive relatives from Swiss-Prot and Trembl. That is why we put the column "The number of accurate results" into the Table 1. There is a number of sequences, more then 20% similar to their presumptive Swiss-Prot and Trembl relatives.

Prof_Pat families consist of no less then 3 sequences. On the one hand, this kind of relationships is strong and stable and makes result of new sequences examination more reliable. On the other hand, almost one half of Swiss-Prot/Trembl sequences have no 2 or more relatives and therefore have no participation in the Prof_Pat formation. If some unknown

sequence has only one relative protein in the bank, we can't reveal this fact unlike IproClass bank does (Table 1). We plan to correct this deficiency in future.

Conclusion

Prof_Pat was showed to have as good specificity as the best world-known "secondary" banks. At the same time, its completeness and variety of included proteins were higher than those of other banks, and its search speed was 3-10 times higher than search speed of any other protein family bank we examined.

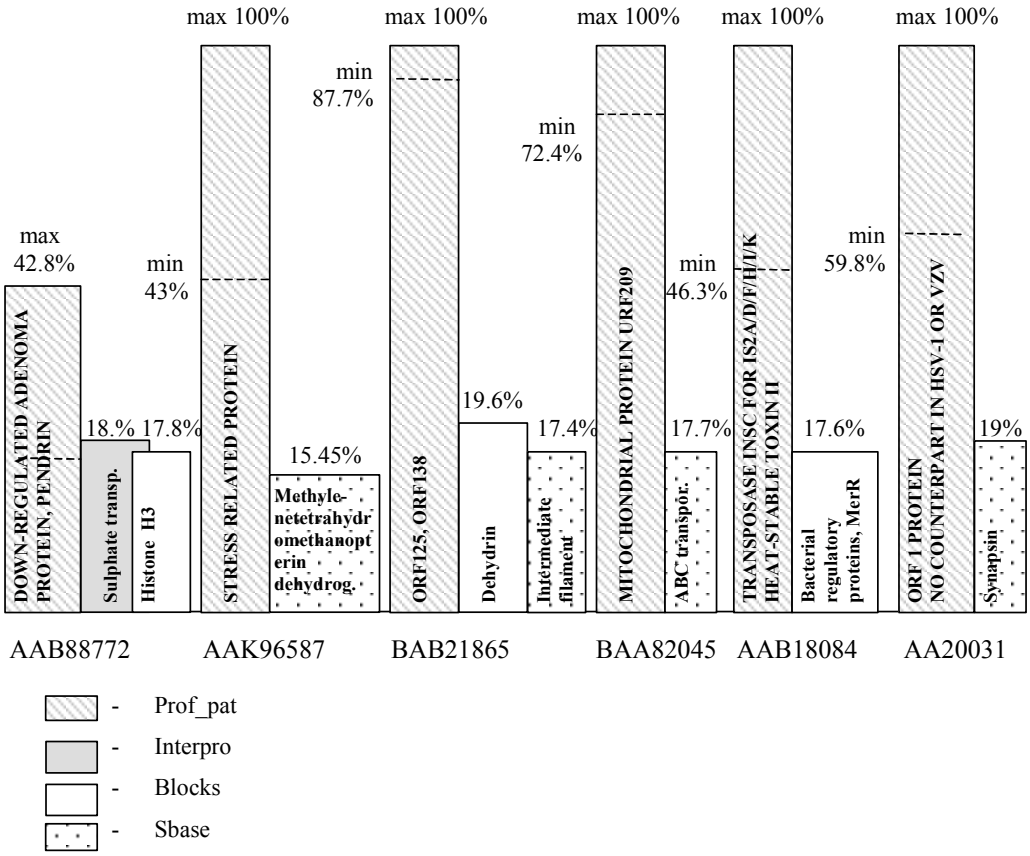


Fig. 1. Some cases of disagreement in the data of Prof_Pat and other banks.

The maximal and minimal similarity level is showed for Prof_Pat and only maximal – for other banks, when sequences were compared to Prof_Pat families and SwissProt/Trembl proteins directly.

References

1. Bachinsky A.G., Yarigin A.A., Guseva E.H., Kulichkov V.A., Nizolenko L.Ph. (1997). A bank of protein family patterns for rapid identification of possible functions of amino acid sequences. CABIOS. 13, 115-122.
2. Bachinsky A.G., Frolov A.N., Naumochkin A.N., Nizolenko L.Ph., Yarigin A.A. (2000) PROF_PAT 1.3: updated database of patterns used to detect local similarities. Bioinformatics. 16, 358-366.

MODELING OF CD150 CYTOPLASMIC TAIL INTERACTIONS WITH SH2D1A AND Fyn SH2-DOMAIN

**** Palagina G.S., Sidorenko S.P.***

R.E.Kavetsky Institute of Experimental Pathology, Oncology and Radiobiology of National Academy of Sciences of Ukraine, Kiev, Ukraine
e-mail: aelytha@yahoo.com
*Corresponding author

Key words: *CD150, SH2-domain, Fyn, SH2D1A, threading, ITSM-motif*

Resume

Motivation: Cell surface receptor CD150 is involved in modulation of cell fate. How CD150 transmits signals is not known. The purpose of this study was the investigation of interactions of CD150 cytoplasmic tail with the components of signal transduction pathways by means of computational biology methods.

Results: Modeling of the CD150 immunoreceptor cytoplasmic tail (CD150ct) was done using threading methods. Docking of the modeled structure with Fyn protein tyrosine kinase Src-homology 2 (SH2) domain revealed residues in the distal immunoreceptor tyrosine-based switch motif (ITSM) of CD150ct, that may be responsible for Fyn/CD150ct interactions.

Introduction

CD150 is a cell surface phosphoglycoprotein involved in cell fate modulation in the immune system. CD150 is found on T-cells, B-lymphocytes, monocytes, macrophages and dendritic cells and can transmit both positive and negative signals, depending on the signaling molecules, which bind the receptor's cytoplasmic tail. It belongs to the CD2 subfamily of immunoglobulin superfamily and shares homology with CD244, CD48, CD84, and CD229 (Shlapatska et al., 2001). Cytoplasmic tail of CD150 contains paired tyrosine-based motif TxYxxV/I within its sequence, which were shown to serve as binding sites for certain SH2-containing proteins after phosphorylation. This motif is called immunoreceptor tyrosine-based switch motif (ITSM). In T- and B-cells, CD150 is associated with protein tyrosine kinases Fyn, Lyn and Lck, adapter protein SH2D1A, tyrosine phosphatase SHP-2 and inositol phosphatase SHIP (Mikhalap et al., 1999; Howie et al., 2002, Latour et al., 2001, Wu et al., 2001). The purpose of this study was to investigate interactions of SH2D1A adapter protein and protein tyrosine kinase Fyn with CD150 cytoplasmic tail by means of computational biology methods.

Methods

Threading of the CD150 sequence. The structural models of the CD150 full and truncated forms were constructed by threading method THOM2 (threading onion model 2) at Lloop server (www.tc.cornell.edu/CBIO/Loopp). The matrices for modeling were selected on the basis of low potential energy and high THOM2 Z-scores. Modeling was performed using the 3D structure of α-cobratoxin from *Naja naja siamensis* (PDB ID 2ctx), which has a CD59 like fold according to CATH classification of proteins (Betzel et al., 1991). CD150 secondary structure prediction was carried out using PHD program (www.maple.bioc.columbia.edu/phd).

Geometry optimization. Further modeling was performed by HyperChem 6.01 program (www.hypercube.org). Energy minimization was carried out with molecular mechanics method. The simulations were carried out in MM+ force field, using Pollack-Ribiere geometry optimization algorithm.

Molecular docking. We used crystal structures of SH2D1A in complex with CD150 peptides, containing the sequence around the first ITSM (Y281) and its phosphorylated versions (pY281) (Poy et al., 1999) for docking of the CD150ct to SH2D1A and Fyn SH2 domain in complex with phosphotyrosile peptide crystal structure (Mulhern et al., 1997; Arold et al., 2001), for docking with Fyn. Docking was carried out with FTDock program from 3D-Dock 3.0 program suite.

Results and Discussion

Sequence-structure alignment of CD150ct and α-cobratoxin made with THOM2 method is shown in Figure 1. Cobratoxin belongs to the SCOP class of the small proteins and is comprised of 71 a.a. residues compared to CD150ct's 77 (Betzel et al., 1991). CD150ct secondary structure was predicted using PHD program (www.maple.columbia.edu/phd).

```
Cd150ct LRRRGKTNHYQTTVEKKSLTIYAQVQKPGPLQKKLDSFPA 1 - 59
2ctx    IRCFITPDITSKDCPN--GHVCYTKTWCDAFCSI-RGKRV 1 - 54

Cd150ct QDPCTTIVVAATEPVPESVQETNSITVYASVTLPES     60 - 76
2ctx    DLGCAATCPTVKT--GVDIQCCSTDNCNPFPTRKRP     55 - 71
```

Fig. 1. Sequence-structure alignment of the CD150ct with α-cobratoxin, obtained at LOOPP threading server by using THOM2 threading method.

Model building. To clarify the mechanism of binding of the CD150ct to SH2-containing molecules, we have built the 3D-model of CD150ct using the 3D structure of cobratoxin (PDB ID 2ctx) found with THOM2 threading method. The amino acid side chains of the matrix were changed into CD150ct residues according to the sequence-structure alignment generated by the LOOPP server. Geometry optimization was performed by HyperChem 6.01 program and it took 598 cycles of Pollack-Ribiere Conjugate Gradient. Optimization was run in vacuo using MM+ forcefield. The energy of the final structure is –465.01 kJ/mol compared to starting parameter of 315.335 kJ/mol. The validity of the obtained structure was measured using WHATCHECK, ERRAT and Verify3D programs (www.doe-mbi.ucla.org/SERVICES). According to our model, the CD150ct has the CD59-like fold, containing three antiparallel β-strands.

CD150ct interaction with Fyn SH2 domain and SH2D1A. Molecular docking was done with FTDock program from 3D-Dock suite. Angular deviation of 15 degrees and electrostatic filtering were applied for docking. To filter out the best complexes we put the following filtration criteria: either pY307 or pY327 were chosen to be located at hydrogen bond distance from Fyn SH2 domain. Fyn SH2 domain was not able to bind the area surrounding pY307, but was able to interact with the distal ITSM (pY327). Analysis of three best complex predictions has shown that residues R19, V59, N95, and H94 in Fyn SH2 domain were critical for this interaction. According to our model, R19 forms hydrogen bonds with phosphogroup of pY327 (Fig. 2), and H94 imidasole ring interacts with S324's hydroxyl group (Fig. 3). These interactions are stabilized with interbackbone contacts shown in Figure 2. Since Fyn's SH2 domain was unable to dock the ITSM motif around pY307, it is possible that this ITSM represents a binding site for other SH2-containing protein tyrosine kinases, for example Lyn in B cells) or Lck (in T cells).

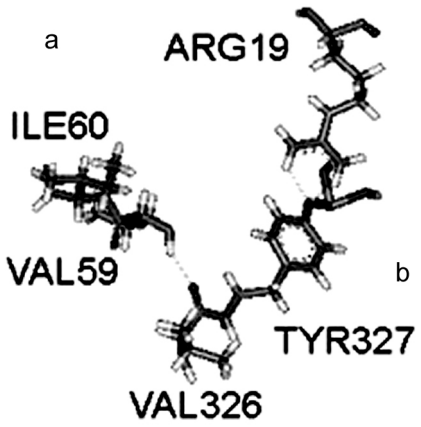


Fig. 2. Interaction of CD150ct distal tyrosine Y327 with Fyn SH2 ARG19 (hydrogen bonds) and interbackbone contacts formation between CD150 VAL326 and Fyn VAL59. a – Fyn SH2 domain, b – CD150ct.

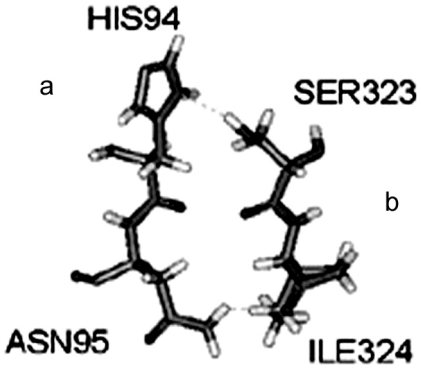


Fig. 3. Interactions of CD150 residues SER323 and ILE324 with residues HIS94 and ASN95, respectively, from Fyn SH2 domain. a – Fyn SH2 domain; b – CD150ct.

Our results support the model of CD150 interactions with SH2D1A and PTK's, according to which SH2D1A's binding to Y281 (pY281) recruits a tyrosine kinase, which phosphorylates Y307 and Y327, thus, creating binding sites for Fyn (pY327) and, possibly, for Lck (pY307).

Acknowledgements

The work was supported by CRDF grant UB-531.

References

1. Arold S.T., Ulmer T.S., Mulhern T.D., Werner J.M., Ladbury J.E., Campbell I.D., Noble M.E.M. (2001) The Role of the Src Homology 3-Src Homology 2 Interface in the Regulation of Src Kinases. J. Biol. Chem. 276, 17199–17206.
2. Betzel C., Lange G., Pal G.P., Wilson K.S., Maelicke A., Saenger W. (1991) The refined crystal structure of alpha-cobratoxin from Naja naja siamensis at 2.4-A resolution. J. Biol. Chem. 266, 21530–21536.
3. Howie D., Simmaro M., Sayos J., Sancho J., Terhorst C. (2002) molecular dissection of the signaling and costimulatory functions of CD150 (SLAM): CD150/SAP binding and CD150-mediated costimulation. Blood. 99, 957-965.

4. Latour S., Gish G., Helgason C.D., Humphries R.K., Pawson T., Veillette A. (2001) Regulation of SLAM-mediated signal transduction by SAP, the X-linked lymphoproliferative gene product. Nature Immunology. 2, 681-690.
5. Mikhalap S.V., Shlapatska L.M., Berdova A.G., Law C.L., Clark E.A., Sidorenko S.P. (1999) CDw150 associates with Src-homology 2-containing inositol phosphatase and modulates CD95-mediated apoptosis. J. Immunol. 162, 5719-5725.
6. Mulhern T.D., Shaw G.L., Morton C.J., Day A.J., Campbell I.D. (1997) The SH2 domain from the tyrosine kinase Fyn in complex with a phosphotyrosyl peptide reveals insights into domain stability and binding specificity. Structure. 5, 1313–1321.
7. Poy F., Yaffe M.B., Sayos J., Saxena K., Morra M., Sumegi J., Cantley L.C., Terhorst C., Eck M.J. (1999) Crystal Structures of the XLP Protein Sap Reveal a Class of Sh2 Domains with Extended, Phosphotyrosine-Independent Sequence Recognition. Mol. Cell. 4, 555–562.
8. Shlapatska L.M., Mikhalap S.V., Berdova A.G., Zelensky O.M., Yun T.J., Nichols K.E., Clark E.A., Sidorenko S.P. (2001) CD150 association with either the SH2-containing inositol phosphatase or the SH2-containing protein tyrosine phosphatase is regulated by the adaptor protein SH2D1A. J. Immunol. 166, 5480-5487.
9. Wu C., Nguyen K.B., Pien G.C., Ninghai W., Gullo C., Howie D., Sosa M.R., Edwards M.J., Borrow P., Satoskar A.R., Sharpe A.H., Biron C.A., Terhorst C. (2001) SAP controls T cell responses to virus and terminal differentiation of Th2 cells. Nature Immunology. 2, 410-414.

A MODIFIED GENETIC ALGORITHM WITH LOCAL AND GLOBAL SEARCH TECHNIQUES

[1,2] Yang Z.L., [1,3] Liu G.R., [3,4] Lam K.Y.*

[1] Center for Advanced Computations in Engineering Science (ACES) Department of Mechanical Engineering, National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260, e-mail: mpeliugr@nus.edu.sg
[2] Singapore-MIT Alliance (SMA), E4-4-10, National University of Singapore, 4 Engineering Drive 3, Singapore 117576
[3] SMA Fellow, Singapore-MIT Alliance (SMA), E4-4-10, National University of Singapore, 4 Engineering Drive 3, Singapore 117576
[4] Institute of High Performance Computing (IHPC), Singapore 118261
[*]Corresponding author

Key words: *modified genetic algorithm, local and global search techniques, performance test, speed of convergence*

Resume

Motivation: Genetic algorithms are important searching tools in bioinformatics computing. The problem of slow in convergence is one of the key problems of using genetic algorithms in bioinformatics. Improvements are needed to be done for these genetic algorithms and make them efficient enough.

Results: A one-step Pattern-move direct searching technique is used in the local search procedure to improve the performance of a standard genetic algorithm. In the global search procedure, a pre-treatment is made to ensure the random selection of individuals to be those that did not occur previously. The results of performance of testing functions have shown great success in speeding up the searching procedures by using the local and global searching techniques. Besides, as there is no derivative to be used in this approach, great forward calculation saving can be obtained comparing to other gradient methods. The modified genetic algorithm can be used in the computation of bioinformatics just the same way as other standard genetic algorithms.

Availability: http://www.nus.edu.sg/ACES/

Introduction

The first major work on genetic algorithms was carried out by John Holland (Holland, 1994) as a method of searching for global optimum in complex systems. The concept of genetic algorithm comes from the principles of Charles Darwin's theory of biological evolution (Michalewicz, 1994). According to Darwin's theory, during the development of species, individuals compete with each other in order to survive. The strong individuals are able to survive and have more offspring because they are better suited to the environment. After generations, an increasing number of individuals inherit from the strong ones, while genes from weaker individuals trend to disappear gradually. Besides, offspring from strong parents may even be better suitable to the environment than their parents. As time goes on, the species becomes better adapted to the environment.

In many cases, the solving procedures of optimization problems share similar principles with those of how species develop. In these cases, the procedures of finding a desire solution can be regarded as a process evolving towards perfect adaptation to an environment. The potential solutions can be regarded as individuals, while the objective function of the problem is the environment. The desire solution is the one that can lead to the best value of objective function. Genetic algorithms utilize the ideas in searching algorithms by letting a population of candidate solutions (individuals) evolve towards a true solution of a problem. According to the algorithm, the individuals (with several genes) are coded into binary bits and the individuals of each generation are produced through three stochastic operations: selection, crossover and mutation. Primitive genes of individuals are selected randomly according to the genetic operators. The crossover operator recombines a pair of individuals into two new individuals. The mutation operator alters a single individual by change its genes (binary bits). Genes are decoded again into variables for evaluation. The stronger individuals (solutions with better fitness) are selected for reproduction and the weaker individuals (solutions with bad fitness) are replaced by the stronger ones in the new generations.

As random selection is one of the three genetic operators, this feature makes genetic algorithm a very robust algorithm in finding the global optimum rather than local optimum in multi-optimum problems. This advantage is very important in the problems of prediction of protein structures (Patton et al., 1995; Dandekar et al., 1994; Unger et al., 1993), RNA structures (Ogata et al., 1995; Titov et al., 2000; Shapiro et al., 1994) and so on (Isokawa et al., 1996; Deaven et al., 1995; Wayama et al., 1995). However, because of the random selection, the time required to find the desired solution is usually very long. The searching time will also increase very rapidly as the number of genes in the individuals increase. It is commonly believed that the genetic algorithm is impractical for finding the global optimum for real life problems with large number of genes, unless measures are taken to speed up the searching process.

In this paper, first, a modified genetic algorithm with local and global search techniques is introduced. The performance is then tested using several testing functions, comparison of speed of convergence with standard genetic algorithm is finally made.

Modified genetic Algorithm

To speed up the convergence procedure, in the following part, the standard genetic algorithm is modified using local and global searching techniques.

In the local searching procedure, a one-step Pattern-move direct search strategy is adapted. Pattern-move is a direct search strategy, which can be used to find out the best point around the current points. The combination use of Pattern-move local searching technique and genetic algorithm is performed as following: Suppose p_j and p_{j-1} are the two best individuals in the j^{th} and $(j-1)^{th}$ generation, respectively. In order to find out better individuals around the best individuals in the j^{th} generation, two new individuals c_1 and c_2 are generated through the forward and internal interpolations, respectively. This can be expressed in the following equations:

$$c_1 = p_j + \gamma_1(p_j - p_{j-1}) \quad (1)$$

$$c_2 = p_j - \gamma_2(p_j - p_{j-1}) \quad (2)$$

Where γ_1 and γ_2 are two nonnegative decimals whose values can be changed to adjust the distances between these new individuals and original individuals p_j and p_{j-1}. To get stable convergence, generally, the ranges of these parameters are: [0, 1.0]. For simplicity, in the rest of this paper, we fix the values of γ_1 and γ_2 to be 0.2 and 0.5, respectively.

In the global search procedure, a pre-treatment procedure is made to ensure random selection of individuals to be those that did not occur previously. This can be done through defining a vector to remember the individuals that are used. Therefore, the domain of candidate individuals for random searching in this method becomes progressively smaller as the searching goes on.

Results of performance test and discussion

In order to compare the performance of present method with standard ones, the selected testing functions are those typical functions used for performance testing. They are listed in Table 1. Table 2 gives the results of testing functions in Table 1, performance comparison between genetic algorithm with local searching technique and the standard one is performed, the percentage rates of number of generations for desirable fitness of present method over standard one are listed in this Table.

Table 1. Test functions used in the performance test.

F1	$f(x) = \sin(x) + \sin(10x/3) + \ln(x) - 0.84x + 3, \quad 2.7<x<7.5$
F2	$f(x_1, x_2) = \prod_{i=1}^{2} \sin(5.1\pi x_i + 0.5)^6 \exp\frac{-4\log 2(x_i - 0.0667)^2}{0.64}, \quad 0 < x_i < 1.0$
F3	$f(x_1, x_2) = \sum_{i=1}^{5} i\cos((i+1)x_1 + i) * \sum_{i=1}^{5} i\cos((i+1)x_2 + i) + \left((x_1 + 1.42513)^2 + (x_2 + 0.80032)^2\right), -10 < x_i < 10$
F4	$f(x_1, x_2, x_3) = \sum_{i=1}^{3}\left(\left(x_1 - x_i^2\right)^2 + (x_i - 1)^2\right), \quad -5 < x_i < 5$
F5	$f(x_1, x_2, x_3) = \sum_{i=1}^{3}((ax_1 - bx_i^2)^2 + (cx_i - d)^2)$, where, $a = 0.99934, b = 1.00056, c = 0.99904,$ $d = 1.00094$, $-5 < x_i < 5$
F6	$f(x_1, x_2, x_3) = \sum_{i=1}^{10}\left[e^{(-ix_1/10)} - e^{(-ix_2/10)} - \left(e^{(-i/10)} - e^{(-i)}\right)x_3\right]^2, \quad -5 < x_i < 15$
F7	$f(x_1, x_2, x_3, x_4) = \sum_{i=1}^{5}\frac{1}{\sum_{j=1}^{4}(x_j - d(i,j))^2 + c(i)}, \quad 0 < x_i < 10$

Table 2. Generations for convergence using genetic algorithm (GA) with local searching algorithm and standard GA.

No	Test function	Modified GA		Standard GA		Ratio
	$f_{optimum}$	N_M	f_M	N_c	f_c	N_M/N_c
F1	-1.601	7	-1.601	313	-1.601	2.24%
F2	1.0	110	1.0	10868	1.0	1.01%
F3	-186.7	219	-186.7	11363	-186.7	1.93%
F4	0.0	409	-2.235E-8	34618	-2.235E-8	1.18%
F5	0.0	306	1.232E-5	26487	1.023E-4	<1.1%
F6	0.0	730	-1.459E-8	39997	-.917E-7	<1.8%
F7	-10.15	250	-10.15	20576	-5.101	<1.2%

Table 3. The number of forward calculations between genetic algorithms with and without global searching technique (GST).

Test Function	GA with local search technique			Standard GA		
	Without GST	With GST	Saving (%)	Without GST	With GST	Saving(%)
F1	805	357	55.65	2500	1236	50.56
F2	3980	2116	46.83	100000	56490	43.51
F3	33305	13368	59.86	75000	43147	42.47
F4	42260	19627	53.56	200000	140639	29.68
F5	40220	22075	45.11	150000	105496	29.7
F6	8050	4859	39.64	200000	135878	32.06
F7	8005	5195	35.1	250000	190692	23.72

The comparison of the number of forward calculation between genetic algorithms with and without global searching algorithm is made in Table 3. The percentage saving of number of forward calculation is also listed in this Table.

The comparison of the number of forward calculation between genetic algorithms with and without global searching algorithm is made in Table 3. The percentage saving of number of forward calculation is also listed in this Table. The typical searching procedures using present modified GA and standard GA are shown in Fig.

Compare to standard genetic algorithm, only a small fraction of generations is needed to get desired solution using this local searching algorithm (Table 2). Furthermore, if the global search technique is used, the number of forward calculation can be further reduced (Table 3). The improvement can be even greater if the combination approach of local and global search techniques is used. Besides, the present algorithm keeps all the advantages of standard genetic algorithms and can be easily used in the same way as the standard ones. These show the great effectiveness of the proposed method.

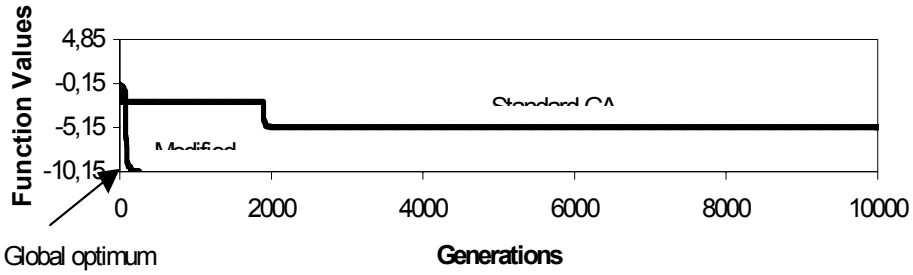
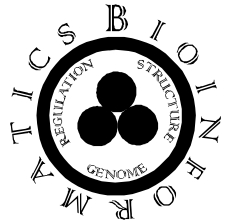


Fig. Searching procedures of test function F7 in table 1 using present modified GA and standard GA.

References

1. Dandekar T., Argos P. (1994) Folding the Main Chain of Small Proteins with the Genetic Algorithm. J. Mol. Bio. 236, 844-861.
2. Deaven D., Ho K. (1995) Molecular geometry optimization with a genetic algorithm. Phys. Rev. Lett. 75, 288-291.
3. Holland J. (1994) Adaptation in Natural and Artificial Systems, MIT Press.
4. Isokawa M., Wayama M., Shimizu T. (1996) Multiple sequence alignment using a genetic algorithm. Genome Informatics. 7, 176-177.
5. Michalewicz Z. (1994) Genetic algorithm + data structures = evolution programs, Springer-Verlag, New York.
6. Ogata H., Akiyama Y., Kanehisa M. (1995) A genetic algorithm based molecular modeling technique for RNA stem-loop structures. Nucl. Acids Res. 23, 419-426.
7. Patton A., Punch W., Goodman E. (1995) A Standard GA Approach to Native Protein Conformation Prediction. Proc. Sixth Intern. Conf. Gen. Algo. (ed. L.Eshelman) 574, Morgan Kaufmann .
8. Shapiro B.A., Navetta J. (1994) A massively parallel genetic algorithm for RNA secondary structure prediction. J. Supercomputing. 8, 195-207.
9. Titov I.I., Ivanisenko V.A., Kolchanov N.A. (2000) Comp. Tech. 5, 48.
10. Unger R., Moult J. (1993) Genetic Algorithms for Protein Folding Simulations. J. Mol. Bio. 231, 75-81
11. Wayama M., Takahashi K., Shimizu T. (1995) An approach to amino acid sequence alignment using a genetic algorithm. Genome Informatics. 6, 122-123.



METHODOLOGICAL PROBLEMS OF BIOINFORMATICS

EFFICIENT METHODS FOR ADEQUATE GRAPHICAL PRESENTING MOLECULES AND MOLECULAR COMPLEXES

Kravatsky Y.V., * Nikitin A.M.

Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia, e-mail: amnikitin@eimb.ru
*Corresponding author

Key words: *molecular graphics, molecular viewer*

Resume

Motivation: Interactive presenting and manipulating molecule images is an essential step in modeling and constructing of molecules. Available tools of this kind are slow and the resulting image quality is low even for molecules composed of 1 000 atoms; moreover, most of them are operating on workstation class computers. Our work is an attempt to solve this problem on the basis of common personal computers.

Results: We have developed a set of methods for adequate presentation and interactive manipulation of molecules and molecular complexes, which was implemented as a computer program. For instance, our viewer program produces a quality images and allows interactive manipulation (rotation, zoom, etc.) of a model representing a molecular complex of 10 000 atoms without annoying delays even at Pentium II-class PC.

Availability: Viewer program for Windows 9x/NT is available on request from the authors.

Introduction

One of the most important problems both in molecular biophysics and in biochemistry is to display obtained results correctly and interactively. It is an essential step of any scientific process, because in many experiments and/or calculations final result cannot be clearly presented as one number, or a few of numbers, so result can be vague from the first point of view. Graphical presentation is extremely useful in numerous cases when the number of parameters, which can describe the object of interest, is very large.

Here we will speak only of small part of great problem of really fast but correct presentation of gathered scientific data. We will discuss the problems arising when computer graphics is applied for molecular biophysics, especially when tasks of displaying and operation with visual images of molecules and molecular complexes are being solved.

Unfortunately, we have to contend that during last decade or even more there have been no noticeable progress in development of methods intended for displaying and treatment of scientific 3D data. All programs and projects we know grow rather extensively than intensively. As the result, performance of programs, which are created for displaying, rotating etc. molecule images, is still low even at SGI platform and hardly allow to operate interactively with molecules more than 1 000 of atoms in the CPK representation (*Insight II, Discover*), nevertheless modern PCs outperform ten-years-old workstations by many times (right now 1 GHz PC with 256M RAM is rather common than something unique).

The way to solve this problem is to use for the scientific programs algorithms and approaches, which are developed by the 3D computer games developers. Usually, at the every step of data processing, researcher has no need in photorealistic or publishing quality images. He (or she) has need rather in interactive program that, of course, can display data adequately, from the correct point of view, but with no abusive graphical and calculational embellishments like multiple sources of colored light, accurate shading, surface texture implementing etc.

Contemporary 3D molecular program has to be a kind of visual meccano realized in the integrated development environment (IDE) for chemists and molecular biologists. In this IDE researcher can has capability both investigate molecules and assemble them from the blocks and easily change their geometry by means of mouse as well as numerically. Advanced investigator is to have possibility to define his own building blocks. Most of existing systems lacks this feature. Other weak point of software already available is that most packages are designed for protein investigations while modern software should has equally good capabilities to work with proteins, small molecules and DNA (*HyperChem*). Due to current great interest to genome investigation, serious DNA displaying and modelling support has to be implemented in any modern molecular graphics software.

There is one problem, which is technical rather than scientific, but which is very important for all researchers. It is required for contemporary molecular designing software system to be cross- platform, or, at least, it has to be PC-oriented (*RasMol, RasTop*). Traditionally, complex scientific calculations and graphical packages was implemented under the UNIX clones solely (*Insight II, Discover*). But right now almost all laboratories all over the world have personal computers, which performance comparable with UNIX workstations. There is no serious molecular visualising and modelling chemical software for the PCs right now.

Methods and Algorithms

Having in mind all the above details we shall try to build the contemporary software system for display and investigation of large molecules and molecular complexes. First of all 3D graphics concept of the future package should be defined. We believe that two light sources with no color for adequate molecule representation should suffice. Moreover, shading calculation or side position of the light source (in our case the light source can move only along Z-axis) is not essential for understanding the molecule structure.

The viewer of the molecular visual model uses the following techniques improving the speed of picture rendering without affecting the image fidelity.

1. One of the most CPU time-consuming problems in the picture rendering is approaching the full-screen resolution, especially in the case of molecule with relatively large number of atoms (over 1 000). In this case it can be useful to decrease the screen resolution intentionally while the molecule is moved or rotated by a researcher. Multiplicity of the picture resolution downgrade depends on the CPU power and is automatically calculated from a simple test. The real picture resolution can be decreased from full-screen to 2, 4, 8, 16 or even 32 times to save interactive manner of operation with molecule. If CPU power is still insufficient, the program ignores the hardly observable atoms (represented by just a few pixels).

During idle time the viewer improves the picture resolution (up to full-screen). Figure 1 presents the Hoechst 33258 molecule fragment (bis-benzimidazol) in the draft (a) and final full-screen (b) modes. Comparison of these two pictures demonstrates the common structure and peculiarities of molecule even in the draft variant.

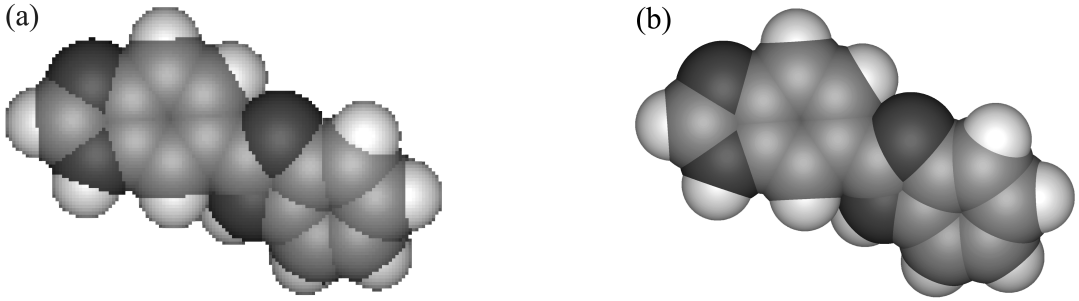


Fig. 1. Hoechst 33258 molecule fragment in the preliminary resolution (a) and in the final (b) resolution mode.

2. Z-buffer is believed to be the most efficient algorithm of hidden surface removal (Rogers, 1985) and we used this technique in the described viewer.

3. Picture rendering can be accelerated by the atom sorting along the Z-axis in the direction from the observer to the screen in order to exclude the atom parts hidden by the front atoms. It proves especially useful when a large molecule is oriented more or less along Z-axis.

4. Further improvement of the above method includes additional sophisticated filters. The first one completely removes surely hidden whole atoms from further consideration. The second one considers the visible atom and calculates its hidden segments (to 1/8 accuracy) in order to remove them from further consideration.

5. Complex lighting model with the numerous colored light sources is not essential for the attractive molecule picture. As was already mentioned, it is possible to build fine molecule image with two light sources—point white located at the Z axis in the observer's eye and a diffused one. Atom images obtained with this lighting model are centrally symmetric. The algorithm that calculates the surface color and brightness in the general case is an n^2–proportional method while in our centrally symmetrical case it is an n-proportional method (n is the number of iterations).

6. Integer-point calculations and function tabulation wherever possible can further accelerate the rendering.

Implementation and Results

Figure 2 shows the fragment of 10859 atoms B-form DNA image generated by the above techniques. This molecule can be interactively manipulated in the described viewer without annoying time delays even at PC Pentium II-350; unfortunately it cannot be confirmed by a static figure.

Another useful feature of the molecule display program is clipping the front part of molecule or molecular complex in order to see the hidden internal parts of it. This feature is illustrated in Fig. 3 demonstrating the presence of a unoccupied space inside the buckminsterfulleren C60 molecule.

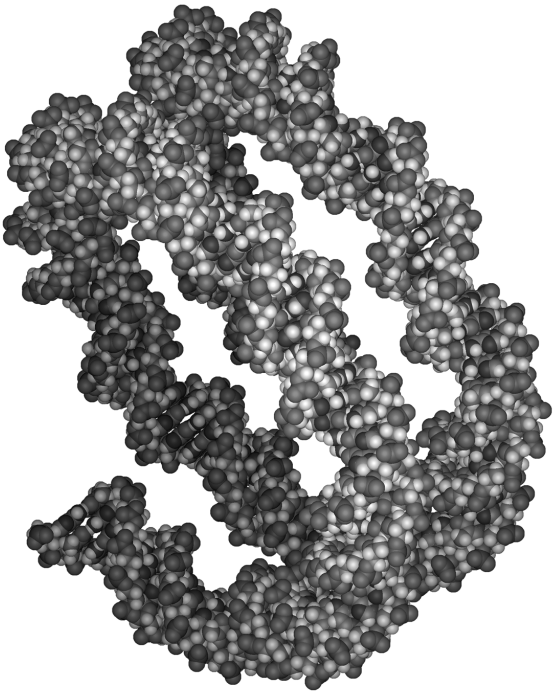


Fig. 2. 10859 atoms B-DNA fragment.

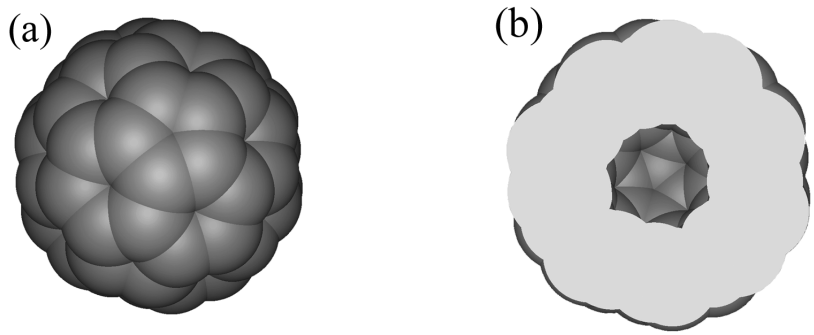


Fig. 3. Buckminsterfulleren C60 molecule external (a) and clipped (b) view.

Discussion

Future development of the described viewer have to be done as its modification to the system that both displays molecules adequately and performs various calculations for them. This system has to be a CAD system clone for biochemists and molecular biologists and must have a comprehensive set of best available computational methods. The system designed for a serious researcher should allow its easy extension by user. It can be realized by supporting a scripting language and possibility to link user's program modules with the system.

Recent success in genome decoding greatly increases the interest in this field. The huge volume of the obtained information increases the significance of automated investigation methods and, hence, to the comprehensive systems which can perform adequate molecular investigations.

References

1. http://www.hyper.com, HyperChem.
2. http://www.msi.com, Insight II, Discover.
3. http://www.openrasmol.org, RasMol, RasTop.
4. David F. Rogers (1985) Procedural Elements for Computer Graphics. McGraw-Hill Book Company, New York.

NATURAL CLASSIFICATION OF NUCLEOTIDE SEQUENCES

[1] Vityaev E.E., [2] Kostin V.S., [3] Podkolodny N.L., [4] Kolchanov N.A.

[1] Institute of Mathematics, SB RAS, Novosibirsk, Russia
[2] Institute of Economics and Industrial Engineering, SB RAS, Novosibirsk, Russia
[3] Institute of Computer Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
[4] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
e-mail: vityaev@bionet.nsc.ru
*Corresponding author

Key words: *bioinformatics, knowledge discovery and data mining, machine learning, eukaryotic promoter recognition, transcription factor binding sites*

Resume

Motivation: A principally new approach to constructing classifications of nucleotide sequences on the basis of the "natural" classification concept is proposed in the paper. The "natural" classification is based on the following principle: objects of one class should obey one group of rules, and objects of different classes should obey different groups of rules. Based on this principle, a method for constructing a classification, an algorithm, and a GeneNatClass software system have been developed.

Results: A method for constructing the classification, algorithm, and the GeneNatClass software system have been developed, which allows identification of "natural" classes of subsequences, that is, motives.

Availability: Scientific Discovery Website: http://www.math.nsc.ru/LBRT/logic/vityaev

Introduction

Numerous principles of constructing classifications are currently known. The classifications are based on the hypothesis of compactness and various measures of closeness in a certain space, on resemblance of standards, on supertargets, on various criteria of classification quality and quality functionals, on separation of distribution mixtures, etc. (Classification, Clustering, 1977).

Nevertheless, these approaches rarely yield stable and law-like results. Therefore, they should be used carefully, with clear understanding of restrictions in conclusions that can be drawn on the basis of these classifications.

In contrast to the above-listed classifications, the objective of a "natural" classification is the insight into the object domain and identification of those latent relations, notions, and regular features that are important for constructing the theory of the object domain and possessing a predictive force. In this meaning, "a natural classification is only the one that reflects the law of nature" and ensures the following (Zabrodin, 1981; Vityaev, Kostin, 1992):

Prediction of the maximum of object properties, based on the object place in the classification;

Maximum of general statements about each class;

Retaining of the structure of classes with variation of classification features; and

Objectivity, reliability, and predictive force.

A constructional criterion of a "natural" classification was proposed in (Vityaev, 1983): "Objects should be divided into classes with accordance with the rules satisfied by the objects. More exactly, objects of one class should obey one group of rules, and objects of different classes should obey different groups of rules. Objects of one class should also possess some integrity, which is understood as mutual agreement of rules of each group in terms of mutual prediction of object properties. In addition, groups of rules may have common features that establish relations between features of objects from different classes".

Sets of rules of each class reveal a regular structure of objects of the class. A regular structure exactly reflects the idealization process (Vityaev, Kostin, 1992); therefore, the structure itself is called an ideal object of a class, and the classification procedure is called idealization.

The method of "natural" classification (Vityaev, Kostin, 1992) may be divided into the following stages:

1. Mapping of raw data and formation of a learning sample.

a) Formalization of various types of relations important for the description of chosen objects from the viewpoint of an expert.

b) Constructing a feature space of objects. Constructing higher-order variables from other primitive variables.

2. Data cleaning and preprocessing. Constructing data samples.

3. Finding rules.

a) Specification of a system of nested Rule Types.

b) Generation of various statistical hypotheses on the basis of Rule Types and their verification on the learning sample; search for rules relevant for recognition of various types of objects.

3. Search for all regular structures (ideal objects) of classes.

Methods and Algorithms

Nucleotide sequences are used as initial data. To construct a learning sample, one has to define a specification of objects and their properties. A set of features measured on these objects determines various relations between nucleotides, their positions, sections of sequences or full nucleotide sequences, etc.

In the general case, the set of features whose values are determined for particular objects may change. Formally, these data can be represented in the form of an XML description or a set of relational tables.

Algorithm for finding rules. To find rules, we use the relational approach to Data Mining methods (Kovalerchuk, Vityaev, 2000) verified by the Discovery system, which allows one to find and test almost all types of hypotheses in the first-order language. A system of nested Rule Types implements a strategy of more and more detailed and exact analysis of the object domain theory. This approach allows one (1) to process data of all types, measured in arbitrary scales (partial-order, grids, titles, orders, log-linear, trees, networks, graphs, etc., and also mixtures of these quantities) and (2) to find law-like rules under conditions of noise and small learning samples.

The simplest rules on symbolic sequences have the form

IF $(Pos_{i1}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon 1}\&...\&(Pos_{ik}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon k}$

THEN $(Pos_{i0}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon 0}$, (1)

where $(Pos_{ij}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon j}$ that one of the values of {A|T|G|C} is located (for $\varepsilon_j = 1$) or is not located (at $\varepsilon_j = 0$) in the position ij of the object **α**. We denote the hypothesis of rule (1) after the condition IF as Premis(**£**).

Statement (1) is a rule if the following conditions are satisfied:

$\text{Prob}(\text{Premis}(\boldsymbol{£})) > 0;$ (2)

$\text{Prob}((Pos_{i0}(\boldsymbol{\alpha})=\{A|T|G|C\})^{\varepsilon 0})/\text{Premis}(\boldsymbol{£}))>\text{Prob}((Pos_{i0}(\boldsymbol{\alpha})=\{A|T|G|C\})^{\varepsilon 0})/\text{SubPremis}(\boldsymbol{£})$.

Here, Prob is the probability of the statement and SubPremis(**£**) means that one or several relations are lacking in the hypothesis. If the second condition of (2) is not satisfied, then, if some relation from the hypothesis Premis(**£**) is deleted, the conditional probability of some subrule **IF** SubPremis(**£**), **THEN** $(Pos_{i0}(\boldsymbol{\alpha}) = \{A|T|G|C\})^{\varepsilon 0}$) is not lower than the conditional probability of the rule itself; hence, this relation can be deleted.

The relational approach to Data Mining methods means the use of a strategy of a more and more exact and detailed analysis of data by means of arbitrarily complicated refinement of the hypothesis being tested. For instance, hypotheses for symbolic sequences may include additional features of belonging of nucleotides to a certain region, specific or admissible distances between nucleotides, properties of nucleotides themselves, etc.

All features of objects are tested as target features of the rule. The hypothesis (Premis) plays the role of a filtering query that chooses those objects from the sample for which all features of the hypothesis have the values indicated in the rule. To measure the rule force, we compare the conditional distribution of target values obtained when all hypothesis features are satisfied and the distribution of target values on all objects. The stronger the rule, the greater the deviation of the conditional distribution of the target values from the initial distribution. One of the simplest methods of measuring this deviation is the statistics χ^2. We use it in the form of the so-called normalized $Z\chi^2$-deviation:

$$Z_{\chi^2_{i_0}} = \sqrt{2\chi^2} - \sqrt{5}$$

$$\chi^2_{i_0} = \sum_{k=1}^{2}\sum_{j_0=1}^{4}\frac{(N_{kj_0} - E_{kj_0})^2}{E_{kj_0}}$$

N is the total number of rows in the table;

N_k is the number of rows in the table where the hypothesis (k=1) is satisfied and the hypothesis (k=2) is not satisfied;

N_{j0} is the number of rows in the table where the values of the target feature $j_0 \in \{ATGC\}$ are observed;

$E_{kj0} = N_k N_{j0}/N$ is the expected value of N_{kj0} under the condition that k and j_0 are independent;

N_{kj0} is the number of rows in the table where the values k and j_0 are observed simultaneously.

The probability inequalities (2) are tested by this $Z\chi^2$-deviation. The rules are sought by gradual increasing of the rule hypothesis by one feature at each step. The extended hypothesis should yield a stronger rule than the short one.

Construction of ideal objects. The next stage of analysis of nucleotide sequences is the construction of ideal images of real-world sequences. If the objects of a class possess some integrity, it is manifested in the structure of regular relations unifying the parts/features of an object into a single unit. It is the structure of regular relations that determines the unification of the parts/features of an object into a single unit.

The idealization procedure reduces to the following. Using all rules, we supplement the description of a real object by additional values of features that are predicted with a high probability by the remaining set of features and already included features and delete those features that fall outside the overall ensemble. This procedure is continued until all the necessary values are included and all random values are sorted out. This procedure is regulated by the criterion of mutual agreement of rules, which should be rigorously increasing at each step.

In terms of software, the idealization process is implemented as follows: for a certain real sequence, a matrix M is generated, which contains the number of rows equal to the number of nucleotides in the sequence and four columns (one for A, T, G, and C). The entire set of rules R is considered. Each rule applied to the sequence adds its four predictions in the form of $Z\chi^2$-deviations (for A, T, G, and C) into four cells of the row corresponding to the target feature of the rule. If the sequence contains one or more of these four values, the total criterion of self-consistency acquires a contribution equal to the sum of the predictions ($Z\chi^2$-deviations) of these values (and these values only). If the sequence contains the value with the negative sign, the corresponding contribution is also taken with the minus sign. The occurrence with the negative sign means that the sequence should not contain the corresponding nucleotide. Zero indicates that this nucleotide does not enter this sequence, but the absence of this nucleotide is not required.

We determine the sequences that are **ideal objects of classes.** For this purpose, we introduce a criterion of mutual agreement of the rules on prediction on these objects:

$$\Gamma(\mathrm{M}) = \sum_{R} \sum_{j_0=1}^{4} Z_{\chi^2_{i_0}} \delta(i_0, j_0)$$

where R is the set of rules and $\delta(i_0,j_0) = 1$ (-1) if the state of the nucleotide $j_0 = \{A|T|G|C\}$ in the current position i_0 of the sequence coincides (does not coincide in the case of –1) with the values in the sequence itself.

DEFINITION (Vityaev, 1983). **An ideal object** of a class is the set of nucleotides <{A|T|G|C}{A|T|G|C}…{A|T|G|C}> for which the criterion Γ(M) has a local maximum (the value of this criterion rigorously decreases in the case of deletion or addition of an arbitrary value in this set). The record {A|T|G|C} means that one or several nucleotides from those indicated in the braces can be treated as an ideal object.

Conclusions

A software system GeneNatClass has been developed. The system implements the above approach to constructing "natural" classifications. The system developed has been tested on problems of classification of splicing sites and transcription factor binding sites; the efficiency of the system has been demonstrated.

Acknowledgements

The authors are grateful to I.B.Rogozin for providing splicing site data. This work was partly supported by the Russian Foundation for Basic Research (Grant № 01-07-90376, 00-04-49229, and 02-07-90355) and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Vityaev E.E., Kostin V.S. (1992). Natural Classification as the law of Nature. In: Intelligent systems and Methodology. Proc. Symp. "Intelligent supporting of activity in complex subject domains", Novosibirsk, 7-9 Apr., 1992, part 4, Novosibirsk, 1992, 107-115 (In Russ.).
2. Vityaev E.E. (1983). Classification as a determination of groups of objects that satisfy different sets of consistent regularities. Computational Systems. 99:44-50 (In Russ.).
3. Zabrodin V.Yu. (1981) Criteria of naturalness of classifications. NTI, ser. 2. Classification and Clustering, Ed. J.Van Ryzin, Academic Press, New York, 1977.
4. Kovalerchuk B., Vityaev E. (2000). Data Mining in Finance: Advances in Relational and Hybrid methods. Kluwer Academic Publishers, 308.
5. Vityaev E.E., Orlov Yu.L., Vishnevsky O.V., Kovalerchuk B.K., Belenok A.S., Podkolodny A.S., Kolchanov N.A. (2001). Computer System "Gene Discovery" for Functional Annotation of DNA Sequences. Proc. Workshop Machine Learning as Experimental Philosophy of Science, ECML/PKDD 2001 (Eds. K.B.Korb, H.Bensusan), Freiburg, Germany, 3-7 September, 2001, Freiburg University, 1-11.

LOGICAL SPECIFICATION OF NEURAL NETWORKS

**** Mikhienko E.V., Goncharov S.S., Vityaev E.E.,***

Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia, e-mail: nord@land81.nsu.ru
*Corresponding author

Key words: *bioinformatics, Neural Networks, Knowledge Discovery, Machine Learning, semantic probability inference, MofN rule, logical programs, specification*

Resume

Motivation: Though neural networks are extensively used to tackle the problems associated with bioinformatics, they are still a "black box" the outcome of which cannot be interpreted. To solve the problem, a neural network specification should be created that could serve to interpret the result from the standpoint of a problem area.

Results: Based on MofN rules proposed by Shavlik, a logical representation of a neural network is constructed with the subsequent creation of neural network specification as a logical program. To realize the possibilities of the specification obtained, comparison of neural networks and semantic probability inference is made that forms the basis of computer system Gene Discovery designed in the Institute of Cytology and Genetics, SB RAS. The prediction ability of knowledge obtained by semantic probability inference is shown to be better than that of knowledge provided by neural networks.

Availability: Scientific Discovery Website: http://www.math.nsc.ru/LBRT/logic/vityaev

Introduction

The concept of a neural network was first introduced by McKalloc and Pitts in 1943. Since than neural networks have evolved unevenly, with ups and downs in the research. Nowadays neural networks are a widely used and generally accepted as data analysis method for bioinformatics. Experiments show that neural networks give better results than conventional techniques. At present, a mathematical formalism of neural networks is well developed, and a great body of information on their application in various fields of knowledge is accumulated. However, many problems and restrictions remain to be dealt with. Principal problems with neural networks are as follows.

1. The lack of neural network basis. Though described by mathematical terms and formulae, neural network construction and training are of pronounced empirical nature. There exists no well-defined algorithm to construct the structure of a neural network solving one or another problem; this is achieved by numerous experiments.

2. No assurance that the network obtained will be optimal.

3. A neural network is still a "black box" the outcome of which cannot be interpreted in the concepts of the object domain.

4. Despite all their advantages, the currently available neural networks fail to investigate the object domain, and cannot serve as a research tool, i.e., the problem of knowledge extraction from a network exists.

Actually, all basic problems are related to reliability and understanding of the result obtained by neural network. It is not sufficient to arrive at some result; one should know how it was derived by the network as well as the degree of its reliability. As long as there is no possibility of interpreting the results of neural networks, they cannot be widely used in the field where accuracy is the main criterion, and errors cause severe consequences. Moreover, in practice, in any field new methods are necessarily subjected to expert examination, therefore, their practical implementation depends on the possibility of network interpretation. We believe that further progress in the application of neural networks will become possible when the above problems are resolved in one or another way. One of the possible variants of the problem solution is to indicate the method giving such a representation of a neural network that will allow one to perform the necessary analysis and estimate the results obtained by the network.

In fact, we mean the construction of a formal specification of a neural network. To do this, one has to specify the language or formal system wherein the problem is posed and solved. As is shown in paper [3], formal systems for problem statement and solution must be weak. In particular, logical programming is just such a system. Therefore, the method describing neural network in the language of logical programs may be used. Such a representation is both conveniently analyzed and compared with other techniques of new knowledge extraction.

Methods and Algorithms

At the moment only the papers of J. Shavlik present rather rigorous method for the extraction of rules from neural networks. The method relies on KBANN and MofN algorithms. KBANN algorithm is employed to construct a neural network for the investigation of a problem area using the available knowledge of it. MofN algorithm is applied to extract rules of a definite form accessible for analysis from the network. When used jointly on the basis of the knowledge already available and neural network techniques, these algorithms provide new knowledge thus enriching the Knowledge Base. The approach

proposed by J Shavlik has shown very good results in solving the problems of gene sequence recognition. Mention should be made of papers [4], [5] wherein J Shavlik's methods serve to solve bioinformatics problems.

Such an approach is not the only machine learning method for the extraction of new knowledge from data. Another method to be considered is a semantic probability inference developed by E.E.Vityaev and based on semantic programming theory devised by Yu.L.Ershov, S.S.Goncharov, and L.D.Sviridenko in 1985. The idea of semantic programming is that a computational process should be treated as a statement truth test using some model. Moreover, one can study a variety of relations between statements and model considering computational process as definition of probability or statistical significance. The reason for the development of semantic probability inference is that the currently existing methods of new knowledge extraction make use of axiomatic approach that fails to give adequate estimates of the reliability of the knowledge obtained. Essentially, the problem is as follows: in axiomatic approach estimates of decisions have no effect on the inference process, thus they can decrease considerably in the course of logical inference. As a result, the decision has an uncertain degree of reliability, and it is not clear in what sense it is a decision. Semantic probability approach reveals the facts from which the decision follows with the maximum value of conditional probability estimate. Probability inference calls for no inference rules, and is well determined by the increase in the probability estimate. Computational process may be represented as a tree, with the objective A $\leftarrow$ at the vertex, and the rules derived by adding an atom or a conjunction of atoms to the premise – at the nodes. The choice of the atom to be added is governed by the requirement for the conditional probability to increase. So we have a conditional identity the premise of which involves a conjunction of atoms belonging to the set of data of the problem area under consideration. The conditional probability of the resulting conditional identity is greater than that of any other conditional identity the premise of which includes atoms belonging to the same set of data, and the inference – the objective A. Estimates of probabilities and conditional probabilities are calculated by the probability logic rules. Computer system Gene Discovery based on semantic probability inference served to reveal new regularities in DNA sequence, and analyze the sequence of erythroid-specific promoters and endocrine system ones [1, 2].

Results

Let us describe the logical specification of neural networks. Let us show that any MofN rule can be represented as a set of conditional identities. General form of the rule MofN is as follows: if m of $\{A_1,\ldots,A_n\}$ is valid, then A_0. Consider a neural network and the set {MofN} of rules describing it. Denote the atom that is j neuron of i layer by A_j^i, then the rule MofN takes the form: if m of $\{A_{j_1}^i,\ldots,A_{j_n}^i\}$ is valid, then A_j^{i+1}. The set of conditional identities corresponding to this rule consists of C_n^m formulae the premise of which contains the conjunction of different m atoms of the set $\{A_{j_1}^i,\ldots,A_{j_n}^i\}$. Denote the conjunction of atoms that is the conditional identity premise by $\{A_{j_m}^i\}$. Thus conditional identities representing the rule MofN are of the form $\{A_{j_m}^i\}\rightarrow A_j^{i+1}$. Neurons of the first layer are represented by facts $A_j^1\leftarrow$.

Example. The rule "IF 2 of $\{A_1^2,A_2^2,A_3^2\}$ is valid, THEN A_1^3 " is represented by the following set of $C_3^2=3$ formulae $A_1^2\,\&\,A_2^2\rightarrow A_1^3$; $A_1^2\,\&\,A_3^2\rightarrow A_1^3$; $A_1^2\,\&\,A_2^2\rightarrow A_1^3$. The set of conditional identities with added facts that describes all rules MofN representing a neural network is called a logical program. Exactly this logical program will be a logical specification of the neural network. Denote the program representing the network as $\Pr_{NN}(A)$, and the program obtained by semantic probability inference – as PR(A). Then the theorem is valid:

Theorem. If atom A is predicted by the program $\Pr_{NN}(A)$ from data D with the estimate $\eta_{NN}(A)$, then it is predicted by the program PR(A) from the same data with the estimate $\eta_p(A)$ such that $\eta_{NN}(A)\le\eta_p(A)$.

References

1. Vityaev E.E., OrlovYu.L., Vishnevsky O.V., Kovalerchuk B.K., Belenok A.S., Podkolodnii N.L., Kolchanov N.A. Computer system "Gene Discovery" for functional annotation of DNA sequences. ECML'2001, Freiburg, Sept., 2001. 1-11.
2. Vityaev E.E. Semantic Approach to Knowledge Base Creation. Semantic Probability Inference. Vych. Sist., 146, Novosibirsk, 1992, 19-49 (In Russ.).
3. Ershov Yu.L., Samokhvalov K.F. On a New Approach to the Philosophy of Mathematics. Vych. Sist., 101, Novosibirsk, 1984, 141-148 (In Russ.).
4. Noordewier M.O., Towell G.G., Shavlik J.W. Training knowledge – based neural networks to recognize genes in DNA sequences. IAdvances in Neural Information Processing Sistems (v.3), San Mateo, CA: Morgan Kaufmann, 1991.
5. Maclin R., Shavlik J. Using Knowledge-based Neural Networks To Improve AI-gorithms: Refining the Chou-Fasman Algorithm for Protein Folding. Machine Learning, 11, 1993, 195-215.

GENOTYPE CLASSIFICATION AND ALLELIC PATTERN RECOGNITION USING KOHONEN SELF-ORGANIZING MAPS

[1] Yuryev A., [2] Makeyev A.

[1] Orchid Biosciences Inc., Princeton, NJ USA
[2] InforMax Inc, Bethesda, MD USA
[2] e-mail: maa@informaxinc.com

Key words: *phenotype, genotype classification, allelic pattern recognition, Kohonen map, computer analysis, single nucleotide polymorphism*

Motivation

New high-throughput genotyping methods to measure single nucleotide polymorphism (SNP) have been developed. These methods will produce enormous amount of SNP genotyping in order to determine genetic determinants of multi-factorial disease and complex phenotypes. New mathematical methods are required to analyze such volume of data and find genetic patterns in phenotypic clusters.

Introduction

The new algorithm for determination of allelic patterns from a large set of individual genotypes is being developed. The goal for the method is to analyze genotypes as large as 100,000 markers from more than 100 individuals. The method uses Kohonen self-organizing maps for genotype classification based on allelic patterns present in genotypes. If prior phenotypic classification of genotype owners is provided it is possible to evaluate the quality of algorithm classification based on its correlation with phenotypic classes. The special scoring function has been designed to track quality of genotype classification. The function reflects the quality of genotype clusters, their separation and size. If genotype classification is successful the next step is to find allelic pattern, which causes classification.

The allelic pattern is defined as a minimal set of markers required to achieve the same or better quality of genotype classification, which has decreased dispersion relative to the non-clustered genotypes. The markers are sorted by allele dispersion first. The most dispersed markers are removed temporary from all clustered genotypes and self-organizing map classification is repeated. The quality of the second classification is assessed using the scoring function. If classification quality is better or the same as the first one the marker set is removed permanently and next most dispersed set of markers gets removed for evaluation. The elimination of most dispersed markers continues until the clusterization quality starts decreasing. Following parameters and conditions must be investigated for successful algorithm development: classification quality scoring function, optimal genotype and sample sizes, optimal size of marker set used for random elimination and optimal parameters for Kohonen neural network algorithm. The code optimization must be performed for further algorithm acceleration.

Methods and Algorithms

Self-organizing map algorithm. Initial weights for all SOM neurons were identical and equal to $1/\sqrt{9882}$. The first genotype was put on the neural map by choosing the "winner" neuron at random. The "winner" neuron was then trained using linear update function $Wij^{new} = Wij^{old} + \alpha*(Xij - Wij^{old})$ where α is learning rate coefficient ($0< \alpha < 1$) and X is genotype vector. The same updating function was used for entire winner neuron neighborhood. The cell distance was used to define neighborhood size. The optimal initial neighborhood size was found to be 4 for majority of successful clusterization, however some random genotypes clusterized better when neighborhood size was 2 or 3. The next "winner" neuron was calculated for the next genotype to put genotype on the neural map. Its neighborhood was updated again and process continued until all genotype vectors were put on the map. One round of neural map update using all genotype vectors defines the learning "epoch". The size of the updateable neighborhood for winner neuron was decreased linearly with number of epochs in our algorithm. The number of epoch was found to be almost not significant for clusterization and was set to 10 for most calculations.

Genotypes and data structure. CEPH genotypes were obtained from ftp://ftp.cephb.fr/pub/ceph_genotype_db/breakpoint_analysis/mbpa_v2/ceph_bp_genotype/.

Genotypes were formatted as allelic integer vectors containing 9882 alleles from 4941 microsatellite markers. Genotype vectors of all 114 individuals were collected into 114x9882 genotype matrix for routine analysis. Binary genotypes are generated having alleles 0 or 1.

Implementation and Results

The general prove of principal has been achieved using 114 CEPH genotypes. CEPH genotypes consist of 4941 microsatellite multiallelic markers. They are classified in six three-generation families based on information in CEPH database. Kohonen neural network proved to successfully clusterize CEPH genotypes into six clusters without any "knowledge" of the relationship.

The next set of the test data is random computer-generated binary genotypes, simulating SNP genotypes. The subset of random genotypes is generated with random allelic pattern. The sensitivity of the method to the pattern length was determined. Patterns as short as 20 markers (40 alleles) can be successfully clusterized using tuned neural net parameters. Allelic pattern determination is currently being tested.

The proposed method has several advantages compared with other known methods for haplotype and allelic pattern determination. It does not require information about genotypes of relatives and provides researcher with initial assessment of the possibility for genetic determinants and allelic patterns in phenotypic group(s) under investigation. The sample can also be "cleaned up" from noisy genotypes and/or markers to improve clusterization quality using the marker elimination by dispersion approach. The method can be applied to search for genetic determinants of multiloci phenotypes and complex genetic traits.

Discussion

The proposed method shows robust clusterization of both simulated and experimental genotypes. The current efforts are aimed to find the algorithm for allelic pattern determination using the clustering information. The required properties for the method should be robustness and short calculation time. Once the pattern recognition algorithm is created the method can be used by to determine genetic minimal set of genetic markers required for genotype clusterization. This minimal set can be defined as genetic pattern. The pattern can be used for further correlation analysis with known regulatory and biochemical molecular pathways to build the molecular model of diseases and phenotypes.

References

1. Andrade M., Casari G. et al. (1997) Classification of protein families and detection of the determinant residues with an improved self-organizing map. Biol. Cyber. 76, 441-450.
2. Kohonen T. (1995) Self-Organizing Maps, Springer, Berlin.
3. Kohonen T., Hari R. (1999) Trends Neorosci. 22, 305-308.
4. Simula O., Kangas J. (1995) Neral Networks for Chemical Engineers (Bulsari, A.B., Ed.), 371-384, Elsevier Science. Amsterdam.
5. Stahl M., Taroni C., Schneider G. (2000) Mapping of protein surface cavities and prediction of enzyme class by a self-organizing neural network. Protein Engineer. 13(2), 83-88.
6. Toronen P., Kolehmainen M. et al. (1999) Analysis of gene expression data using self-organizing maps. FEBS Let. 451, 142-146.
7. Ultsch A., Siemon H.P. (1990) Proc. INNC'90, Intern. Neral Network Conf., Dordrecht, Netherlands, 305-308.

APPLICATION OF THE METHODS OF INTELLECTUAL DATA ANALYSIS TO SOLVING THE PROBLEMS OF BIOINFORMATICS

[1] Zagoruiko N.G., [1] Pichueva A.G., [1] Kutnenko O.A., [2] Borisova I.A., [2] Kochetov A.V., [2] Ivanisenko V.A., [2] Nikolaev S.V., [2] Likhoshvai V.A., Ratushny A.V., [2] Kolchanov N.A.

[1] Institute of Mathematics, SB RAS, Novosibirsk, Russia, e-mail: [1]zag@math.nsc.ru
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
* Corresponding author

Key words: *taxonomy, choosing of characteristics, recognition, prediction, cleavage site, expression, mutation type*

Resume

Motivation: In the area of molecular biology and genetics a considerable body of experimental data has been accumulated that can be effectively analyzed using the methods of intelectual data analysis and the recognition of implicit empirical patterns.

Results: Application of the methods of data intelectual analysis and the recognition of implicit empirical patterns was demonstrated for solving the three problems of bioinformatics: (1) the prediction of a quantitative level of translational mRNA activity by analyzing the context peculiarities of their functional regions; (2) the recognition of signal peptide processing sites in amino acid sequences; and (3) the identification of the character of mutual impairments in genetic networks by analyzing the curves of dynamics of their variables.

Availability: Description of methods in www.math.nsc.ru/AP/oteks.

Introduction

The methods of intelectual data analysis are used to automatically detect empirical patterns for solving the problems on classification, the recognition of images and prediction [1]-[3]. A peculiarity of these methods is that they are oriented to the problems that can be hardly solved by traditional statistic methods. These are the problems on the analysis of a great body of data, poorly-posed tables (the number of characters being comparable with the number of objects), affected by noise and blanks, with the characters measured on the scales of different types without any grounds for offerring hypotheses on distribution laws, etc.

Many problems of bioinformatics exhibit these peculiarities. Among the most widespread applied problems on data analysis, the problems of the following types can be distinguished [1]-[3]: automatic classification (taxonomy), the choosing of the system of informative characters; the recognition of images, the filling of blanks in data tables.

Below we consider the peculiarities of the problems of these types using three concrete examples corresponding to the different levels of molecular-genetic system (mRNA, protein, genetic network) organization; (1) the prediction of a quantitative level of translational mRNA activity by analyzing the context peculiarities of their functional regions; (2) the recognition of signal peptide processing sites in amino acid sequences; and (3) the recognition of the type of mutation impairments in genetic networks by analyzing the curves of dynamics of their variables. A more comprehensive description of approaches and results is available in the papers of a given issue [5]-[7].

Prediction of the quantitative level of translational mRNA activity: algorithm ZET

It is known that the context characteristics of the functional regions of eucaryotic genes affect their activity during expression [4]. However, the mechanisms of this phenomenon are so far not fully understood. This problem can be effectively solved by revealing the patterns of both the different characteristics of a nucleotide sequence and the level of expression which enables determination of significant parameters and their interrelations. Some significant (i.e., affecting the functional activity of a gene region) characteristics are experimentally available. However, their major part is, probably, unknown. For example, the explicit significant characteristics are the structure of expression signals (the extent to which the most active variants are approached), the presence of negative signals (such as AUG-codons and stable hairpins in 5'NTS). A systematic analysis is necessary for describing these significant characteristics. To this end, we distinguish the different functional regions for yeast genes, i.e., a basal promotor (responsible for transcription), mRNA 5'NTS (initiation of translation), 3'NTS (cytoplasmic matrix stability), and a protein-coding part. As context characteristics, we used the nucleotide composition (mononucleotide frequencies, the ratios between the frequencies of complementary nucleotides, deviations in dinucleotide frequencies).

Thus, the significant context characteristics related to expression intensity are to be determined on a basis of these data. This problem can be solved with the help of algorithm ZET ([1]-[2]), used to predict the values of the elements omitted in the

table of data of the "object-property" type and to "edit" (check) the entire table [1]. Some rebundancy is observed for real data tables as they contain similar objects (lines) and the characters (columns) depending on one another. The ZET algorithm is used to reveal these relations for predicting the value to be found. Each line in the table corresponds, in this case, to a gene and each column corresponds to a certain context characteristic. As a predicted parameter, we used the adaptation index of codons CAI accounting for the level of expression.

The operation of algorithm ZET includes three stages. In the first stage, for a given blank *b(ij)* , from the initial matrix,we choose a "competent" submatrix in the form of *t* lines (objects) most close to the *i*-th line and *t* columns (characteristics) most correlated with the *j*-th column.

In the second stage, the parameters are automatically chosen for the formula used to predict the omitted element. In the third stage, the element is predicted from this formula. The actual prediction error is taken as $d=[b(ij)-b^*(ij)]/\bar{b}(j)$, where *b(ij)* is the true value, $b^*(ij)$ is the predicted value and $\bar{b}(j)$ is the mean of the values of the *j*-th column.

The ZET program was realized in the "editing" regime with a size of a competent submatrix of 3*3. The values of CAI activity were predicted for 171 yeast genes and the expected errors in these predictions were obtained. The problem was solved for each of four functional gene regions separately. The final predicted value was obtained by four variants according to the strategies described in [5]. The results predicted by different strategies were compared by the mean value of the actual prediction error and the greatest correlation between the predicted and true values of CAI parameter. These were the criteria for choosing the strategy that allows for the averaging of predictions from three functional gene regions (best with respect to the expected error). The results obtained using this strategy are shown in Fig. 1 (for details see [5]).

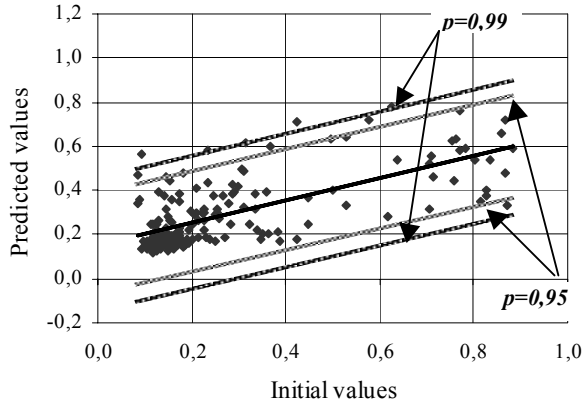


Fig. 1. Confidence intervals of prediction of CAI value with probability p=0,99 and p=0,95.

Thus, the method, based on the ZET algorithm, is proposed for both the prediction of gene expression level from the characteristics of gene functional regions and the combined estimation of the expression level. Besides, the method can be used to draw conclusions on the information density of the separate characteristics of mRNA regions which can testify to the extent to which they affect the expression level. For details aee [5].

Recognition of signal peptide processing sites: algorithm AdDel

The automatic recognition of signal peptides and cleavage sites in proteins is the actual problem of both the recognition of their intracellular localization and the solution of the applied problems of medicine and biotechnology. In [6], the methods of image recognition are considered and the possibility of using the physico-chemical characteristics of amino acids for solving a given problem is studied. The teaching sample is represented by the three sets of eucaryotic protein fragments: the fragments containing cleavage sites ("Signal peptide" image), the fragments containing anchors ("Anchor" image), and the fragments of nuclear and cytoplasmic proteins containing neither sites nor anchors ("Negative" image). In our studies, we used ten Kidera properties and 434 structural and physico-chemical properties of amino acids [6]. As a decisive rule, the " k nearest neighbours" rule was used at k=1.

Seven sets of informative characters were constructed from the given characteristics of amino acids for seven teaching samples. The characters were chosen using the AdDel algorithm [1], combining the ideas of the methods of "sequential addition of the most valuable" (Addition) and "sequential delition of the least valuable" (Delition) characters. For control recognition, the data were imposed taking no part in teaching. The window of a width of *L* symbols moved along the protein chain from left to right with a shift in one symbol and for each window a decision was made on the presence of a cleavage site. The "Signal peptide"/"Negative" images were simultaneously recognized by seven sets of characters and a decision was made by a majority of votes.

In real conditions, of interest are the estimates of the probability of cleavage site presence (absence) in each observation window . These estimates were obtained with the help of a modified "k nearest neighbours" rule using a *membership function* [1]. The technique developed was used to solve the problems on the recognition of the above three images. In this

case, the pairwise comparison method was used [3]. The problem was solved using both the physico-chemical properties and the data on the encounter frequency of amino acids in positions -3, -1 of the analysis window.

The windows of different width and configration were analyzed. It appeared that the best results were obtained for the window with L=8 where the elements with odd numbers are used which is in agreement with the well-known "-3,-1" rule. Figure 2 plots the mean values of the membership function for the window with L=8 of the "Signal peptide" image for 500 control proteins containing cleavage sites. The protein fragments are positioned with respect to the site localization point (vertical line). Figure 3 shows the same for 500 accidentally chosen protein fragments with a length of 70 symbols without cleavage sites.

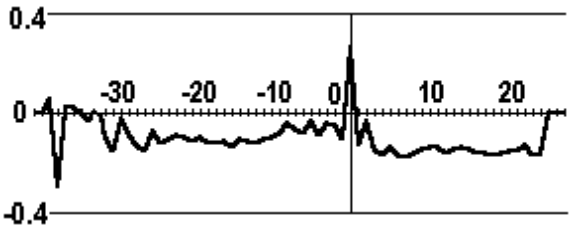


Fig. 2. Membership function for fragments.

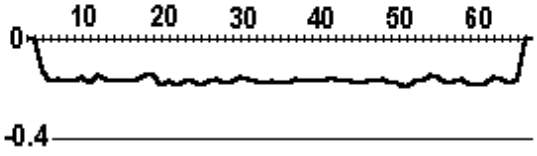


Fig. 3. Membership function for containing cleavage sites for fragments containing no sites.

Experiments on the recognition of three images show a satisfactory discriminating ability of the objects of different images. In this case, taking into account the additional information on the encounter frequency of amino acids in positions -3, -1 of the analyzed fragment increases the recognition quality. Testing a great body of experimental data shows that the cleavage sites are correctly revealed and localized for 85% of cases. For details see [6].

Identification of the character of mutation impairment in genetic networks: algorithm RTM

The recently developed novel experimental techniques (laboratories-on-a-chip) automatically provide kinetic function characteristics in the cells of hundreds and thousands genes and their products. With the availability of such a great body of experimental data on the dynamics of the behavior of the systems under study, the question arises of the technique for their treatment, theoretical analysis and practical use. The actual problem of the analysis of these data is the recognition of the mutation impairments type in genetic networks. Solving this problem will make it possible to develop the methods for diagnosing diseases caused by disturbances in genetic networks operation, the medicines of a strictly specialized action on the given molecular-genetic and biochemical processes occurring in cells, etc.

In [7], a study was made of the genetic network of the differentiation regulation of an erythroid cell under the action of erythroprotein. Using the model of this network, constructed in the framework of the generalized chemico-kinetic method, the data were obtained on changes in the concentration of various substances involved in biochemical reactions. 19 different mutations disturbing the work of a certain link of a genetic network were modelled (for details see [7]) by ten variations for each mutation. The observed kinetics were used as a teaching sample. It was necessary to propose a technique for recognition of the belonging of some cell state to one of 19 mutation types.

The method for solving the problem on mutation type recognition was based on the pairwise recognition principle (for details see [3]). This problem can be solved using the following modification. For each substance and instant of time, its information density was determined by the number of different mutation pairs whose variations from the teaching sample can be correctly distinguished by considering the concentration functions of only this substance and only at this instant. Taking into account these information densities, the approximate algorithm was used to choose the minimum set of substances and for each substance the time intervals of minimum length were chosen to correctly distinguish all variations of all mutations from the teaching sample. The decisive rule was represented as a list in which each pair of compared image standards was correlated with the most informative substance and instants of time.

Recognition was performed using the pairwise delition method [3]. This process was organized so that one of two following results can be obtained in the number of steps not exceeding the number of mutations. We either give the mutation type which is taken as the desired one or claim that the control realization is inconsistent with either of the studied mutation types.

Using algorithm shows that the decisive rule for distinguishing 19 mutation types can be based on the information on the concentrations of only three components of the genetic network: geme (Fig. 4, (1)), receptors bound to transferrine at the surface (Fig. 4, (2)), and mRNA GATA-1 within time intervals of 11, 23 and 2h, respectively. All control mutations were correctly distinguished by the chosen characteristics.

The same mathematical model of genetic network was used to extract data on the genetic network behavior for nine single mutations and 36 double ones forming under the action of all pairs of single mutations. Based on the combination of different proximity measures constructed throughout the entire space of characters, the program provided the list of the most

probable single mutations contained in double mutations (for details see [7]). The correct solutions were obtained for 89% of cases.

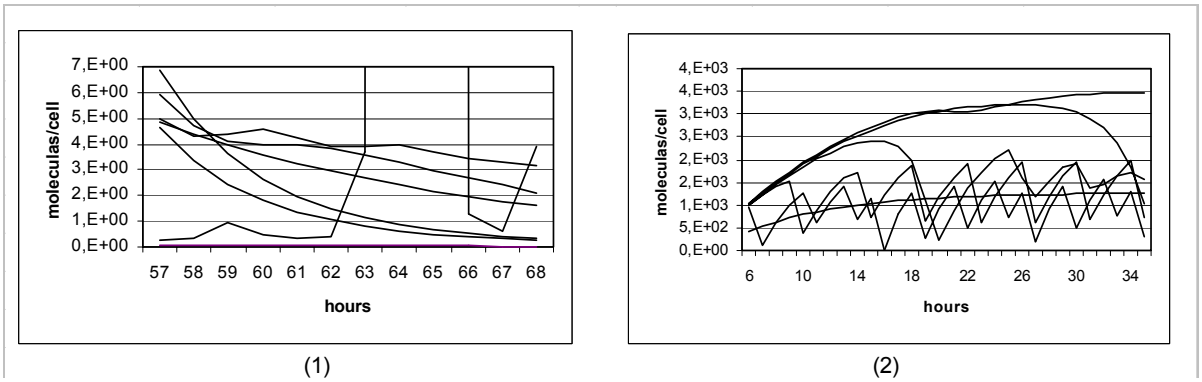


Fig. 4. Dynamics of a change in the concentrations of: (1) geme, (2) receptors bound to transferrine at the surface with different mutations within informative time intervals.

Conclusions

The results obtained testify to the efficiency of the methods of intelectual data analysis in solving a wide range of problems of bioinformatics aimed at the study of the patterns of the structural-functional organization of molecular-genetic systems. The applied approaches are planned to adapt to the peculiarities of the description of initial molecular-genetic data and the sequences of stages of the analysis of several problems typical of bioinformatics.

Acknowledgements

Work was in part supported by the Russian Foundation for Basic Research (grants 01-07-90376, 02-04-48508, 02-07-90355, 00-4-49229, 02-01-00082, 02 01-00082), Ministry of Industry, Science and Technologies of Russian Federation (№ 43.073.1.1.1501), SB RAS (integration project SB RAS № 65), grant of National Institute of Health USA № 2 R01-HG-01539-04A2, grant of the Department of Energy USA № 535228 CFDA 81.049, grant of "Integration", project 274.

References

1. Zagoruiko N.G. (1999) Applied methods of the analysis of data and knowledge. Novosibirsk: izd. IM. 270.
2. Zagoruiko N.G., Elkina V.N., Emelyanov S.B., Lbov G.S. (1986) PPP OTEKS (for data analysis). M.: Finansy i Statistika. 160.
3. Zagoruiko N.G. (2002) Recognition of images by the method of pairwise comparison of standards in competent subspaces of characters. Dokl. AN. M.: Nauka. 382(1), 24-26.
4. Kochetov A.V., Ishchenko I.V., Vorobiev D.G., Kel A.E., Babenko V.N., Kisselev L.L., Kolchanov N.A. Eucaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. FEBS Lett. 1998**,** 440, 351-355.
5. Pichueva A.G., Kochetov A.V., Zagoruiko N.G. A study of interrelations between the level of expression and the context characteristics of functional yeast gene regions by the ZET method. In this volume.
6. Zagoruiko N.G., Kutnenko O.A., Ivanisenko V.A., Nikolaev S.V. Recognition of the presence and localization of cleavage site in signal peptides. In this volume.
7. Borisova I.A., Zagoruiko N.G., Kolchanov N.A., Ratushny A.V. Diagnostics of mutations by analyzing genetic network dynamics. In this volume.
8. Ratushny A.V., Podkolodnaya O.A., Ananko E.A., Likhoshvai V.A. (2000) Mathematical model of erythroid cell differentiation regulation. Proc. of the 2nd Intern Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk. 1, 203-206.

SURVEY OF THE SCIENTIFIC DISCOVERY FOUNDATIONS

[1] Vityaev E.E., Khomitheva I.V.*

[1] Institute of Mathematics, SB RAS, 630090, Novosibirsk, Russia, e-mail: vityaev@bionet.nsc.ru
[*]Corresponding author

Key words: *bioinformatics, scientific discovery, Knowledge Discovery, data mining, KDD&DM, Machine Learning, Goodman's paradox of induction*

Resume

Motivation: During the last decades Knowledge Discovery in Databases and Data Mining (KDD&DM) found considerable applications as systems for extracting certain knowledge from the voluminous, noisy data. The theoretical foundations of KDD&DM are concentrated in Machine Learning and Scientific Discovery. The purpose of the article is to analyze the "induction problem", that is, stating of the problem of induction methods reinforcement for all existing inductive, Machine Learning, Scientific Discovery and KDD&DM methods.

Results: The new stating of induction problem is made in this paper. The paper contains a full and thoughtful survey of Machine Learning and Scientific Discovery methods in a form of a table. The new approach to Data Mining is given. The induction method based on the mathematical model of person's cognitive process has been developed.

Availability: Scientific Discovery http://www.math.nsc.ru/LBRT/logic/vityaev

Introduction

The induction problem is a real and complicated problem of artificial intelligence. Using the data of inductive method reinforces the initial hypothesis in order to obtain a stronger one that gives a more accurate expression to the experiment.

Despite the solid history of the induction problem, still no one has been able to succeed in his (her) attempt in stating the induction problem. This work mainly considers an approach to the induction problem proposed by K.F.Samochwalov (Goncharov et al., 1994), who investigated the possibility of developing induction methods that satisfy some necessary requirements: non-triviality, non-contradictoriness of a hypothesis to initial data, and linguistic invariance (the reinforcement of the hypothesis must not depend on the language it has been formulated). The linguistic invariance requirement uses only such classes of transformations F that hypotheses h and $F*h$ are empirically equivalent, i.e. the observations over the same sets of objects always simultaneously confirm or falsify them. The particular case of linguistic transformations for induction methods is the invariance with respect to the choice of measurement units. However, even these necessary requirements are paradoxical. K.F.Samochwalov obtained the negative result on existence of inductive methods. According to his theorem, there always exists a class of linguistic transformations F that there are no induction methods that are invariant with respect to them. The same theorem was proved in (Vityaev, Novikov, 1988).

Thus, the most general nowadays stating of the induction problem leads to a negative result. So, the correct stating of the induction problem is absent for induction methods from antiquity to nowadays.

Methods and Algorithms

Authors have analyzed the induction problem in order to define the routs of it. In the literature they have found an analog of the negative result—the Goodman's paradox of induction, which also states that the essential requirements are paradoxical. It may be proved that this paradox is a particular case of the linguistic invariance requirement.

Different scientists tried to find a positive way out of the paradox. They arrived to the same opinion that it is more effective not to admit the paradox appearance then to resolve it. For example, von Kutschera (1993) considers that "it is quite possible for several language systems to be equally successful but to lead to different theories about the world" and, thus, "there are no empirical reasons for choosing one language rather than another". Marry Hesse (1969) discusses the problem in such a way: "there can be the cases of alternative sub-theories with equal inductive support, where we do not know which prediction to consider most reasonable."

The authors succeeded in developing the positive interpretation of these results. They came to the conclusion that one can obtain the induction method in the following way:

1) Fixing the language of the induction methods (the types of data, types of hypotheses, and way of manipulating the data);

2) Weakening the linguistic invariance requirement by setting certain limits on the permissible class of transformations.

In order to find such limitations, the authors have examined the mathematical model of person's cognitive process. The choice of a model is not random: human beings proceed effectively in making inductive conclusions through their life. This model enabled the authors to propose the class of homomorphisms entitled the "lower". The theorem has been proved that the class of positive formulas is invariant with respect to the "lower" homomorphisms. Thus, if a class of hypotheses is

fixed as a class of the positive formulas and the a-priori assumption about the "lower" homomorphisms is used, then the regular induction methods exist in the model of person's cognitive process.

Thus, the new stating of induction problem may be formulated as follows: we need to fix the language of induction method, *a priori* assumptions, and class of permissible transformations, then prove the theorem about existing of inductive method. This theoretical way of developing inductive methods has never been used in the history of science.

As initial step to new foundation of inductive methods, we made a full and thoughtful survey of Machine Learning and Scientific Discovery methods in a form of a table. This table states tasks for such areas as Machine Learning and KDD&DM to infer theoretically all methods from the *a priory* assumptions and the language.

In particular, the most general theoretical result was obtained for the class of universally quantified hypotheses. There exist properties of experiment guaranteeing that the hypotheses expressing experimental dependency may be presented as universal formulas. This property of experiment is inheritance of experiment models: if we got experimental result as a model on the set of objects A, then we will get experimental result as a submodel on subset of A. This result states a task of developing an inductive method discovering all the universal formulas as hypotheses.

Let us formulate this method. We will denote as $\nu = \langle P_1,...,P_k \rangle$ the finite set of predicate symbols $P_1^{m_1},...,P_k^{m_k}$, where $P_i^{m_i}$, i= 1,..., k are the m_i-arity predicate symbols; M^ν, the class of all models in ν; pr_0^ν= $\{ D^{\nu,\alpha}(\mathrm{M}) \mid \mathrm{M} \in M^\nu \}$, protocol of experimental results, where $D^{\nu,\alpha}(\mathrm{M})$ is the model diagram; S^ν, axiomatic theory in the first order logic with equality in ν.

The inductive method *I*, which realize the theorem, may be defined as follows $I: \langle S^\nu, pr_0^\nu \rangle \to S_m^\nu$, where S^ν, S_m^ν are the sets of universal formulas in ν and pr_0^ν, training set of experimental results; S^ν is true on pr_0^ν; S_m^ν logically stronger then S^ν.

This inductive method was realized in the program system Discovery (Kovalerchuk, Vityaev, 2000; Vityaev et al., 2001).

This method is rather general and removes all the limitation of the existing KDD&DM methods. Basing on this method, we formulated a new approach to Data Mining methods: Relational Data Mining (Kovalerchuk, Vityaev, 2000; Vityaev et al., 2001).

Let us consider it in more detail. The approach has the following main points:

1) Any Data Mining method assumes explicitly or implicitly defined

a. Data types,

b. Language to manipulate the data, and

c. Hypothesis to be tested on data.

2) Different DM methods ought to consider from the point of view of their Data Types, Languages, and Hypotheses.

3) The program system Discovery has no limitations of the existing KDD&DM methods, because it has the following possibilities:

a) Extension of data types notion using first-order logic with unlimited descriptive possibilities;

b) Using Measurement Theory for describing various Data Types in the first-order logic;

c) Predicate Invention of new features, properties, and Relations defined for Data Types in first-order logic using Measurement Theory.

d) Using any Background Knowledge for Learning and Forecasting.

e) Rule Type notion for describing hypotheses classes; such classes is not defined *a priory* as in the most DM methods; and

f) The notion of law-like rules satisfying all the properties of scientific laws: simplicity, maximum refutability, and logical generality.

The program system Discovery may be considered as a Tool generating a set of forecasting law-like rules by specification of Data Types, Invented predicates, and Rule Types.

The program system Discovery was adapted to gene regulation task in bioinformatics, and computer system Gene Discovery was developed (Vityaev et al., 2001). Thus, the program system Discovery has not only a solid theoretical basis, but also a wide range of applications in various areas (Kovalerchuk, Vityaev, 2000).

In the table, we present only a fragment of the Data Mining methods overview for illustrative purposes. This table states the tasks for all the existing Machine Learning and KDD&DM methods to infer theoretically these methods from the *a priory* assumptions and the underlying language.

General direction in machine learning	Hypothesis space H	Algorithm	The fundamental assumptions of language and data
Concept learning	Each hypothesis $h \in H$ described by a conjunction of constraints on the attributes	Candidate-Elimination algorithm	Language: symbolic or logical representation
Decision tree learning	The hypothesis space is the set of possible decision trees	ID3 algorithm, Sipina	Language: decision tree Inductive bias: "Shorter trees are preferred over longer trees"
Inductive logic programming	Learned hypotheses $h \in H$ are the sets of if-then rules	Foil, Focl	O-priori knowledge: measure of effectiveness Foil_Gain Language: first-order logic rule representation
Analytical learning	$h \in H$ is a Horn clause	Prolog	Domain theory Approximate inductive bias of Prolog: preference for small sets of maximally general Horn clauses

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grant № 02-07-90355) and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Goncharov S.S., Ershov Yu.L., Samochwalov K.F. (1994). Introduction to Logic and Methodology of Science. Interpracs M., Novosibirsk, Institute of mathematics.
2. Vityaev E.E., Novikov V.F. (1988). Induction method paradoxicality. Int. J. Pattern Recognition Artificial Intelligence. 3(1):147-157.
3. von Kutschera F. (1993). Induction and empiricist model of knowledge. Causality.
4. Hesse M. (1969). Ramifications of "grue". British J. Philosophy Sci. 20:13-15.
5. Kovalerchuk B., Vityaev E. (2000). Data Mining in Finance: Advances in Relational and Hybrid Methods, (Kluwer international series in engineering and computer science; SECS 547). Kluwer Academic Publishers.
6. Vityaev E.E., Orlov Yu.L., Vishnevsky O.V., Kovalerchuk B.K., Belenok A.S., Podkolodny N.L., Kolchanov N.A. (2001). Computer system "Gene Discovery" for functional annotation of DNA sequences. Proc. Workshop Machine Learning as Experimental Philosophy of Science, ECML/PKDD 2001, (Eds. K.B.Korb, H.Bensusan) Freiburg, Germany, 3-7 September, Freiburg University, 1-11.
7. http://www.informatik.uni-freiburg.de/~ml/ecmlpkdd/WS-Proceedings/w02/index.html

GENESIS OF THE MECHANISMS UNDERLYING DIRECTED SEARCH FOR BENEFICIAL MUTATIONS

**** Ananko G.G.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: ananko@online.sinor.ru
* Corresponding author

Key words: *rapid evolution, inducer of mutagenesis, sequence hypervariability*

Resume

Motivation: The number of publications devoted to specific mechanisms for generating adaptive variability grows tremendously. A large bulk of experimental evidence devoted to these items and widely spread pattern of their distribution (bacteria, fungi, protozoa, plants, vertebrates) support the fact that these mechanisms could be referred to ordinary genome practice. From the point of view of the evolutional theory, an intriguing task is to determine the conditions and ways how specific mechanisms of adaptive variability are formed. The goal of the current research is to study from the theoretical viewpoint the transition from the random search for beneficial mutations to the search for mutations directed by selection. As an object of studying, we consider a seria of beneficial mutations that consequently appear at a single genome locus. For description of the mechanisms how the local mechanisms of adaptive variability evolve, we have used the algorithm of «random search with self-training».

Results: We have displayed the conditions and mechanism of two-staged selection enhancing local rate of genetic searching for beneficial mutations. This acceleration is achieved due to redistribution of endogenomic inducers of local mutagenesis, without lowering fitness of population.

Introduction

Pathogenous bacteria and trypanosomes are capable to modify rapidly the structure of components that are recognized by host. Specific mechanisms for generating variability were found also in potential hosts of a pathogen, namely, in genes encoding resistance to fungi in plants and in vertebrate immunoglobulins. The rate of generating the variability of key genes responsible for interaction between the organisms determines obviously the results of long-term competition between host and pathogen (Parniske et al., 1997). The high rate of mutational process is conditioned, as a rule, by various endogenomic inducers: mobile elements (Kidwell, Lisch, 1997), various repeats, etc. As shown for some pathogeneous bacteria, distribution of inducers in genome has, on the whole, an adaptive value, that is, it is governed by selection. Also, non-random distribution of trinucleotide repeats that are primarily occurring in the coding parts of genes, was characteristic for genome of Saccharomyces cerevisiae (Young et al., 2000).

One of the useful approaches to studying specific mechanisms of mutagenesis is developed in the works on modeling the hitchhiking effect of a gene-mutator with the beneficial mutation enlarging the total mutational frequency in a genome (Johnson, 1999). However, specific mechanisms of adaptive variability are known to act on many generations without enhancing the average frequency of mutagenesis.

In accordance with the task, we will analyze events in a single genome locus, with participance of endogenomic inducers of mutagenesis with the limited interval of influence (from several dozens nucleotides up to a small group of genes). A series of adaptive cycles, as an object of theoretical study, is useful for revealing some additional properties, which could not be detected in studying the separate cycles (see Discussion).

Algorithm

The algorithm for «random search with self-training» is equally suitable for description of both types of searching, purely random and directed ones. This approach enables to describe the local transition from the random search to the search directed by selection within the frames of a single model. The similar algorithms are used for optimization of complex technological processes (Rastrigin, 1986). The main idea of self-training is that the probability of test searching in profitable directions is enhanced automatically. The model could be viewed as a two-stage process of selection. At the first stage, the perspective genome loci are found, whereas at the second, the intensive search is directed at perspective genome loci «marked» by inducers at the first stage. The criterion of searching for perspective loci is analogous to the simplest heuristic principle: if a single beneficial mutation is present in this locus, then it will be useful to make more close examination of other beneficial mutations. The existence of clusters of potential beneficial mutations in a genome is an essential condition for successful application of this tactics.

An inducer fixed in a definite genome locus is a sort of memory of a population (or a species) about perspective direction of genetical search. The mechanism is optimized by searching for at most possible coincidence of an interval, where an inducer acts (i.e., the size and position of the locus of intensive mutagenesis), to the interval, where a series of potential beneficial mutations are located.

Implementation and Results

Background

All known for today mechanisms generating variability have some features in common. They are constructed from usual genome elements (tandem repeats, mobile elements), localized in particular gene regions. The period of their functioning is comparable to the period of a species existence, that is, their functioning is not accompanied by decrease in population fitness. Specificity is that their action is strictly positioned (in a definite locus or even a site) and that, probably, mutagenesis is restricted by intensity in this case. Selection acts at two levels: at the first one, it favors to increase the fitness of individuals (fixation of beneficial mutation and inducer), at the second, periodical inter-group selection increasing the rate of adaptation in favor of clones (populations), which find novel beneficial mutations faster. Success in the group competition depends upon durability of both stages of adaptation (i.e., the stages of searching and fixation), that is, selection for increase of evolutionary rate should support any mechanisms useful for effective search for beneficial mutations.

Conditions of realization of the mechanism

The mechanism is realized as an auto-regulatory process under combination of the following conditions: 1. The presence of endogenomic inducer, which enhances the frequency of mutations in vicinity of its location. As a local inducer of mutagenesis we denote any element of genome structure, which enlarges the frequency of mutations in particular locus in comparison to the average mutational frequency per genome. 2. The beneficial mutation caused by alteration of positioning of an inducer itself or by DNA variation in the nearest vicinity of the inducer. 3. The presence, under given conditions of inhabitance of population (clone), of the other potential positive mutations within the range of an inducer's action.

The combination of the first two conditions is realized rather frequently due to the fact that most mutations appear under the action of various types of «inner» inducers (Kidwell, Lisch, 1997), that is, the frequency of such combinations is in direct proportion to the share of mutations caused by endogenomic inducers. The third condition is also rather realistic.

Almost all documented cases of rapid adaptive evolution, which are accompanied by series of beneficial mutations, are realized in the variable DNA sequences with the length varying from several base pairs to hundreds of them. As an example may serve a functional divergence of homologous proteins, domains, functional sites, and hypervariable DNA loci of pathogens and their hosts.

Description of a mechanism

At the first stage, the local mechanism of intensive mutagenesis is formed due to direct or accompanying (together with the beneficial mutation) fixation of an inducer (see Fig.). The linked fixation may be of essential importance for adaptive evolution in bacteria and non-sexually reproducing plants. For humans and Drosophila, with recombination rates equaling approximately to 10^{-8}, complete hitchhiking is limited by the interval ranging from 20 to 200 nucleotides (Faya, Wua, 2000).

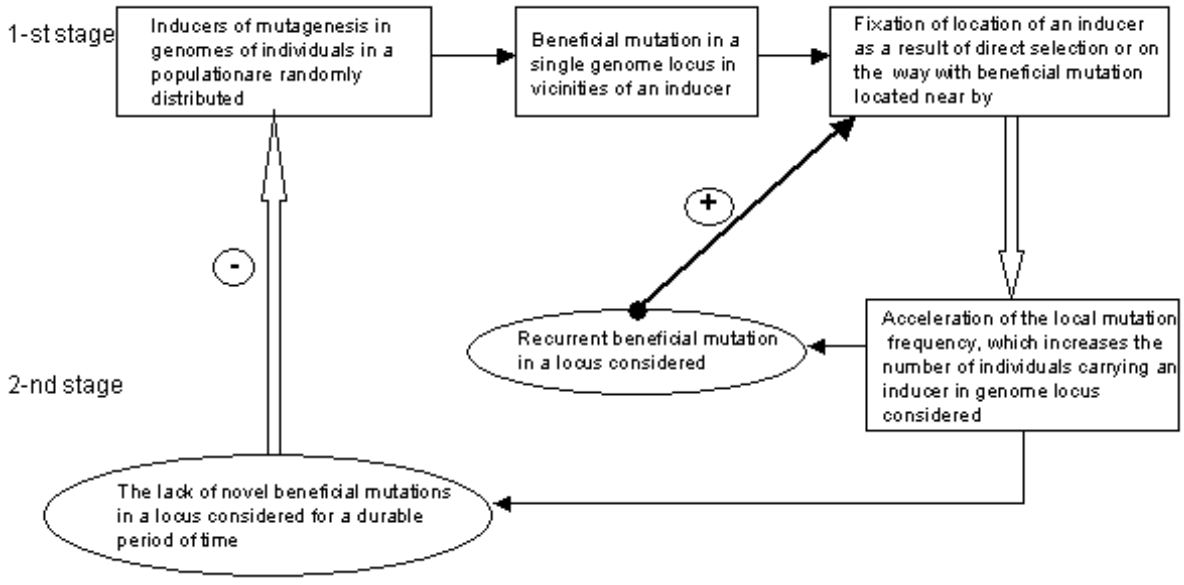


Fig. Autoregulatory mechanism, with positive and negative feedbacks, administering intensity of local mutagenesis.

At the second stage, the mutational events take place that are generated by the inducer fixed in a particular genome locus. According to the properties of an inducer, its localization and the level of fixation in a population, this mechanism may exert strongly varying influence on adaptation. The goal of calculations is to search for potential enhancement of mutagenesis rate due to relocalization of genome inducers. As we analyze long-term processes, then the total number of inducers per population ($\mathbf{N_p*I}$) is considered as a constant value. We account for the following parameters: population size ($\mathbf{N_p}$), number of individuals-carriers of an inducer in a particular locus ($\mathbf{N_f}$), average number of inducers per a single genome (**I**), average number of loci per genome (**L**).

The average number of inducers per a single locus (M_r), for the random distribution of inducers in a genome (i.e., selection is absent) and for the whole population equals to

$$\mathbf{M_r = N_p * I / L.} \quad (1)$$

During fixation, the number of inducers in this particular locus arrives at

$$\mathbf{M_f = N_f + (N_p - N_f) / L + N_p*(I-1)/ L.} \quad (2)$$

For the case of complete fixation, $N_f \rightarrow N_p$, we obtain

$$\mathbf{M_f = N_p + N_p * (I-1) / L.} \quad (3)$$

The coefficient of the relative intensity of induced mutagenesis takes the form

$$\mathbf{E = M_f / M_r = L / I + (I-1) / I \approx L / I.} \quad (4)$$

Thus, the potential of increase in intensity of mutagenesis depends upon the ratio between the number of loci to the number of inducers in a genome. The genome could not be completely saturated by inducers (by definition of an inducer), because for each case, L > I, that is, the frequency of mutations in the locus with the fixed inducer becomes larger than in case these inducers are randomly distributed. Detection of novel beneficial mutations will generate the repeated cycles of fixation of an inducer. In case such beneficial mutations are absent, the system will return to the state with random distribution of inducers (see Fig.).

Besides, in the process of fixation of the first inducer, a probability rises, of fixing the second inducer, which occasionally falls into the locus considered, for many orders of magnitude in comparison to population with random distribution of inducers. In particular, the ratio of probabilities of the triple event (simultaneous occurrence in a particular locus of two inducers and beneficial mutation), under the fixed ($\mathbf{q_{1,2,x}}$) and random ($\mathbf{p_{1,2,x}}$) distribution of the first inducer equals to

$$\mathbf{q_{1,2,x} / p_{1,2,x} = (N_f / N_p) * (L / I*x).} \quad (5)$$

The ratio N_f/N_p increases rapidly and tends to a unit, L > I, while x, a probability of beneficial mutation is an extremely low value. As shown, accumulation of different types of inducers of local mutagenesis really takes place in the genes encoding membrane proteins of gastric pathogen Helicobacter pylori. In this case, the hypervariability of a single gene is supported, as the least, by three different inducers of mutagenesis (Tomb et al., 1997).

Discussion

As the main results of this work, I suggest determining criterion aimed at searching for perspective genome locus and conditions of coincidence of locations of the intervals, where an inducer acts and where a series of potential beneficial mutations is localized. Two-stage selection enables to find perspective loci, thus increasing probability of appearance of next in series members by speeding up mutagenesis in these loci. As shown, in conditions given, the probability of accumulation of inducers in a particular locus grows sharply, that is, intensity of mutagenesis is regulated by varying the number, position of inducers in a locus, as well as the level of their mutability. Hence, the constantly acting, specific mechanisms of the local mutagenesis are possible. This conclusion is supported by already mentioned above non-random distribution of inducers in bacteria and yeast genomes, in integrity with the other phenomenon, predicted by the model, accumulation of various inducers of several types in particular genes of pathogen bacteria.

Except the cases given above, in a series of adaptive cycles, the other relationships could be realized that cause acceleration of adaptation. Since an inducer occupies a similar position in genomes of many individuals in population, one may expect repeated occurrence of similar mutations. Repeated occurrence of beneficial mutation favors to overcome the most difficult period of the stage of fixation, when the random events are prevailing over the influence of selection.

Due to high intensity of mutagenesis in a locus, novel beneficial mutations may appear prior complete fixation of the previous mutation. In this case, some additional effects may appear, in particular, the intensity of selection may rise, whereas the total cost of selection may fall due to linkage of two mutations (Grant, 1985). Thus, the mechanism suggested may not only accelerate the search, but also increase probability of fixation of beneficial mutations.

References

1. Rastrigin L.A. (1986) Happy-go-lucky. M.: Molodaya gvardia. (In Russ.).
2. Faya J.C., Wua C. (2000) Hitchhiking under positive Darwinian selection. Genetics. 155, 1405-1413.
3. Grant V. (1985) The evolutionary process: a critical review of evolutionary theory. New York. Columbia University Press.

4. Johnson T. (1999) Beneficial mutations, hitchhiking and the evolution of mutation rates in sexual populations. Genetics. 151, 1621-1631.
5. Kidwell M.G., Lisch D. (1997) Transposable elements as sources of variation in animals and plants. Proc. Natl Sci. USA. 94, 7704-7711.
6. Parniske M., Hammond-Kosack K.E., Golstein C. et al. (1997) Novel disease resistanse specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. Cell. 91, 821-832.
7. Tomb J.-F., White O., Kerlavage A.R. et al. (1997) The complete genome sequence of the gastric pathogen Helicobacter pylori. Nature. 388, 539–547.
8. Young E.T., Sloan J.S., Van Riper K. (2000) Trinucleotide repeats are clustered in regulatory genes in Saccharomyces cerevisiae. Genetics. 154, 1053-1068.

THE ARCHITECTURE OF CELL DEVICE

[1] ***Tarasov D.S.,*** [1] ***Akberova N.I.,*** [2**] ***Leontiev A.Yu.***

[1] Kazan State University, Russia
[2] Kazan State Academy of Veterinary Medicine, Russia
e-mail: cons@au.ru

Key words*: cell language, cell device, cell simulator*

Resume

Motivation: It's widely accepted that any information processing system can be viewed from two sides: "programmer view' and "engineer view". While engineer deals with physical structure of the system, programmer is interested only in its functional organization that can be relatively independent from system's physical structure. In order to understand any information processing system both points of view must be considered. In this paper we are trying to analyze living cell from programmer's view. Many authors have noticed that the living cell uses some kind of programming language in order to store and express genetic information. This assumes the existence of cell device that reads and executes programs written in the cell language. Understanding the organization of such device could be important for studying regulatory processes in living cell.

Results: based on both biochemical data and formal models of existing human-made computing devices we propose a concept of cell device architecture. This includes the structure of cell device and a hypothetical language used for its programming. A computer simulator of cell device was designed and the compiler was implemented in order to write programs for this simulator. The functionality of proposed model was tested on the examples of well-known genetic regulation systems.

Introduction

Since the first nucleotide sequences begin to emerge, it became clear, that genetic regulation system is rather more complex than it was thought before. Instead of a few monosemantic regulatory sequences, numerous complicated structures have been found, leading to possibility that cell uses some kind of "programming language" in order to store and differentially express genetic information.

Pioneer works in the field of "DNA linguistics" were directed to construction of "dictionaries" of genetic language (Trifonov, Brendel, 1986]. After a large amount of "DNA words" has been collected, the question arose about the way in which they all are linked together. The grammar of genetic language was found to be not context-free and probabilistic models were shown to be inadequate (Collado-Vides, 1991). Many grammatical models were presented, describing the construction rules of different regulatory regions (Collado-Vides, 1992, 1993; Rosenblueth et al., 1996; Leung et al., 2001). Finally, it became evident, that not only DNA, but living cell as a whole, including proteins and other chemical compounds is involved in formation of specific language, called the cell language or *cellese* (Ji, 1999). It was clearly stated that understanding of life requires the application of linguistics.

In order to expand our knowledge, it is useful to look at the living cell from programmer's point of view. In this work we briefly describe our concept of cell device architecture and cell language.

1. Programs, written in cell language contain statements that direct the behavior of intracellular compounds in similar way like computer programs direct behavior of human-made computers.

2. DNA, proteins and other chemical compounds acting together form some kind of "cell device" that can read, execute and modify cell programs.

Here we present our approach to describe the model of "cell device" that takes programs in the special symbolic language with structure proposed to be close to the cell language.

Implementation and Results

The model of abstract cell device (ACD).

All data types and operations allowed in cell device must obey chemical and physical restrictions. This means, that logical structure of cell device should incorporate chemical and physical laws.

The basic concept of cell device model is a metabolic pool – a data space where different types of chemical compounds are stored. All computations in cell device are done by applying transformations to the chemical compounds. Simple chemical compounds (chemical units) are presented by the graph with atoms in vertexes. Only valid chemical structures are allowed

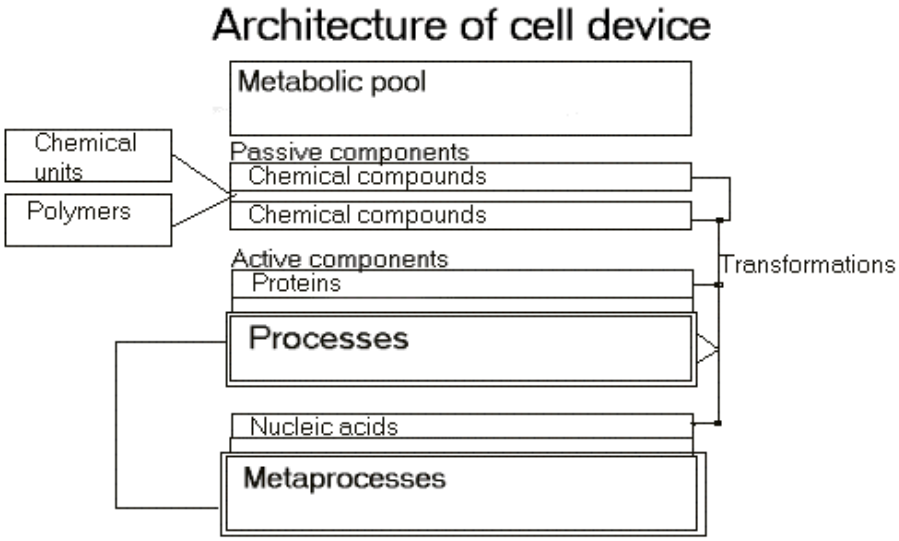
[1] Corresponding author

to be chemical units. In CPDL chemical structure is defined as a type. A metabolic pool can contain one or more instances of chemical unit with defined type. All polymers are handled with in a special way: they are considered as a structure build of several connected chemical units. The polymeric type describes the types of chemical units that can be used to form polymer and allowed connections between them. As a result polymeric units of one type can have different structures.

All chemical compounds are divided into two large groups: active and passive. An active unit can transform other compounds in some predefined way.

One of the important classes of active units is a process. The process has a number of registers and a program that describes its behavior. Chemical units of compatible type can be loaded to the register, undergo some transformations and released. These events occur with probability, defined by internal coefficients and external conditions.

A metaprocess is another class of active units. Its functionality is defined by active processes, available in metabolic pool. Based on set of existing processes metaprocess can generate additional processes.



Implementation of CDS and CPDL Compiler.

We have designed a program that simulates behavior of the cell device and a language for programming such cell device simulator. In order to build cell device simulator and CPDL compiler we must specify a concept of abstract cell device.

Abstract cell device accept only valid chemical structures. By valid chemical structure we mean a structure that can exist in a real world. However it is impossible to define it with the precession enough to implement in CDS. As a result CDS/CPDL compiler may accept illegal chemical structure or (and) deny correct one. We decided to make a set of CDS allowed structures wider than real chemical structures and allow CPDL programmer to define additional restrictions.

Chemically possible transformations. Cell device restricts possible structural transformations to the set of chemically allowed reactions. This problem is close to previous and was solved in a similar way.

Structure and function relation problem. In abstract cell device model the program of active unit is encoded in its chemical structure. But now there is no known way to convert CPDL program to chemical structure. As a result a program of active element has to be associated with structure by the programmer on the basis of experimental evidence.

CDS functional tests.

Two well-known gene regulation systems were used for testing purposes. The lactose operon as the simplest one and the regulation of phage lamda development, as a sample of complex regulatory system. The "computing power" of cell device appears to be enough for correct and full description of these systems in terms of Cell Program Description Language. Furthermore, all CPDL programs, when running on appropriate CDS, show functionality similar to that one could expect for native genetic regulation systems.

Discussion

We assume that most of intracellular processes could be described in terms of the cell device and this assumption is strongly supported by our modelling experiments. CPDL could be used both for qualitative and quantitative modeling (although for the moment only qualitative testing is done).

The main purpose for which CPDL is designed for is to serve as a medium for translation of nucleotide sequence data into the human-readable form, describing corresponding intracellular processes. This problem requires a translation rules from living cell chemical structures to abstract cell device model and as a second step - translation the data of ACD into cell program description language. First task can be easily completed because the structure of abstract cell device is very close to

functional organization of living cell. However, translation from ACD model data to CPDL (and especially back from CPDL to ACD) is a very complicated task due to limitations described above.

References

1. Collado-Vides J. (1991). The search of a grammatical model of gene regulation is formally justified by showing the inadequacy of context-free grammars. Comput. Applic. Biosci. 7, 321-326.
2. Collado-Vides J. (1992) Grammatical model of the regulation of gene expression. Proc. Natl Acad. Sci. USA. 89. 9405-9409.
3. Collado-Vides J (1993) A linguistic representation of the range of transcription initiation of [sigma]^{70} promoters: II. Distinctive features of promoters and their regulatory binding sites. Biosystems. 29, 105-128.
4. Ji. S. (1999) The linguistics of DNA: words, sentences, grammar, phonetics and semantics. Ann. New York Acad. Sci. 870, 411-417.
5. Leung S.-W., Mellish C., Robertson D. (2001) Basic Gene Grammars and DNA-ChartParser for language processing of *Escherichia coli* promoter region. Bioinformatics. 17, 226-238.
6. Trifonov E.N., Brendel V. (1986) GNOMIC: a dictonary of genetic code. Philadelphia: Balaban Publishing.
7. Rosenblueth D., Thieffry D., Huerta A, Salgado H., Collado-Vides J. (1996) Syntactic recognition of regulatory regions in *Escherichia coli*. Comput. Appl. Biosci. 12, 415-422.

ELECTRONIC ENCYCLOPAEDIA IN GENETICS. VERSION 1

[1] Dromashko S.E., [2] Makeyeva E.N., [1] Zheludok A.A*

[1] Institute of Genetics and Cytology, National Academy of Sciences of Belarus, 220072, Minsk, Republic of Belarus
Tel.: +375(17)284-19-44, Fax: +375(17)284-19-17, e-mail: dromash@biobel.bas-net.by
[2] Johns Hopkins University, Baltimore, MD 21218, USA
Tel./Fax: +01-(410) 377-006, e-mail: makeyevaelena@prodigy.net

Key words: *education, electronic encyclopaedia, genetics, information technologies, knowledge bases, molecular genetics*

Resume

Motivation: Genetic dictionaries available in Internet are not numerous and at the best represent HTML pages with cross hypertext links. Russian dictionaries in genetics are not represented at all in Internet.

Results: Russian electronic encyclopaedia in genetics was created based on the book "Genetics: Encyclopaedical Dictionary" by N.A.Kartel, A.M.Makeyeva and A.M.Mezenko (Minsk, 1999). The capacity of electronic version in HTM format is above 30 MB, it is provided with a developed system of hyperlinks and a vocabulary of English-Russian equivalents for the terms interpreted. Electronic encyclopaedia contains 5 thousand glosses, explaining the meaning of terms in classical and molecular genetics, and above 80 illustrations.

Availability: At present electronic encyclopaedia is available only on CD disk.

Introduction

The task of organizing information within the framework of particular fields of knowledge and creation of certain clearly and unambiguously interpreted conceptual base in the form of electronic encyclopaedias and problem-oriented libraries is one of the key tasks among the problems of integrating information resources for scientific and educational purposes. Solution is possible by creating and spreading the corresponding products both on CD disks and via Internet on line mode. Remote access to information resources is more promising since it provides timely renewal of problem-oriented information, its enrichment owing to the results of new investigations.

A need for applying modern information technologies concerns particularly one of the major branches of biological sciences – molecular genetics. Its rapid development in recent decades of XX century resulted in accumulation of a great extent of new knowledge about the essence of life, fundamental genetic processes proceeding in living organism. The level of knowledge achieved in understanding of heredity mechanisms and elaboration of research methods made it possible to proceed to directed engineering of heredity molecules, individual cells and entire organisms. New prospective trends of investigations – genetic engineering and biotechnology arose. Achievements of these sciences are widely used now in medicine, agriculture and industry.

All this has resulted in appearance of a great number of new notions, definitions, terms and symbols earlier not occurred in genetic literature. It is difficult even for a specialist to orientate himself in new molecular-genetic terminology. Besides, sometimes several terms are used definition of one and the same notion in literature or one and the same term has different meanings. Problems associated with true use and interpretation of the meaning of many terms and designations arise often.

So, it is keenly imperative to create new Russian dictionaries of genetic terms and corresponding electronic thesauri and reference- retrieval systems, as well as to integrate them with available databases on molecular-genetic information, for instance, GeneNet (Kolpakov et al., 1998; Kolpakov, Ananko, 1999), for giving users the opportunity to carry out "computer engineering" experiment and "to participate" interactively in the process studied.

Electronic biological dictionaries, available via Internet, first of all in genetics, are not numerous and at the best represent HTML pages with hyperlinks, i.e. they are not equipped with reference-retrieval systems. Besides, most of them are intended for English user. They are Genomics Lexicon (http://genomics.phrma.org/lexicon/) – database developed by Pharmaceutical Research and Manufacturers of America, A Genetics Glossary (http://helios.bto.ed.ac.uk/bto/glossary/index.html) – elaboration of University of Edinburgh, as well as dictionary Glossary of Genetic Terms (http://www.bis.med.jhmi.edu/Dan/DOE/prim6.html), available in site of Human Genome Management Information System, Oak Ridge National Laboratory. Only the dictionary Gentechnik-Lexikon (http://interpharma.ch/info/wissens/lexikon/index.html), compiled by Interpharma, Swiss Pharmaceutical Research Companies, in German is an exception. There are no Russian dictionaries in genetics in Internet.

* Corresponding author

Methods and Algorithms

When creating electronic encyclopaedia we took into consideration our experience of work with databases and WEB design (Panich et al., 1998; Dromashko et al., 2001), as well as the dictionary (Kartel at al., 1999). Standard information technologies and software, usually applied for creating documents in HTM and JPEG formats, were used in this project.

Implementation and Results

As a first step of the project we have created a Russian electronic version of genetic dictionary based on the printed book (Kartel at al., 1999). The remarks made after book publication were corrected in the dictionary, with some terms being defined more exactly with due account of materials that appeared in English literature in 2000–2001.The capacity of electronic version in HTM format is above 30 MB. The dictionary is provided with a vast system of hyperlinks. Electronic encyclopaedia, as a paper version, contains about 5 thousand glosses, explaining the meaning of terms in classical and molecular genetics, and above 80 illustrations. It also includes a vocabulary of English-Russian equivalents of the interpreted terms with hyperlinks to a corresponding gloss in Russian.

On the whole the prepared product fills a gap that existed in the system of Russian problem-oriented electronic dictionaries and may be used in distance education. Unfortunately, the present technical characteristics of BIOBEL server (http://biobel.bas-net.by) of the Biological Science Department at the National Academy of Sciences of Belarus make impossible remote access to the dictionary via Internet. For the present electronic encyclopaedia in genetics is available only in the form of CD disk version.

Discussion

At present there are no similar elaboration in problems of general and molecular genetics in Russian in Internet. The developed information retrieval system (encyclopaedia) will be able to become a model for working out ontology libraries in other fields of information technology applications for needs of genomics, in particular in studying molecular-genetic control systems, gene networks, etc. (Kolpakov et al., 1998; Kolpakov, Ananko, 1999; Kolchanov et al., 2000).

In the future we suppose to collaborate with the Institute of Cytology and Genetics of Siberian Brunch at Russian Academy of Sciences and the Research Engineering Center of Information Technologies at the National Academy of Sciences of Belarus. Such a collaboration gives the possibility to equip the dictionary with developed retrieval system and integrate it with such data- and knowledge bases on molecular genetics as GeneNet (http://wwwmgs.bionet.nsc.ru/mgl/systems/genenet/) and GeneExpress (http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/). Since the above systems are English, it will be necessary to provide encyclopaedia with the problem-oriented system of bilingual Russian-English and English-Russian translation. We plan also to enrich electronic encyclopedia with illustrations using potentialities of advanced video-design and animation. It will give users an opportunity to carry out to some extent "computer engineering" experiment and to participate interactively in processes described in the dictionary.

References

1. Dromashko S.E., Gorbachev A.V., Makeyeva A.M., Panich I.A. (2001) To the elaboration of a scientific and educational system for processing and spreading biological data via Internet. DETECH 2001 – International Workshop on Distance Education Technologies, Maribor, Slovenia, 13–14 September 2001. 203–208.
2. Kartel N.A., Makeyeva A.M., Mezenko A.M. (1999) Genetics: Encyclopaedical Dictionary. Tekhnalogia, Minsk (In Rus.).
3. Kolchanov N.A., Ananko E.A., Kolpakov F.A. et al. (2000) Gene networks. Mol. Biol. (Rus.). 34, (4), 533-544 (In Russ.).
4. Kolpakov F.A., Ananko E.A., Kolesov G.B., Kolchanov N.A. (1998) GeneNet: a database for gene networks and its automated visualization. Bioinformatics. 14, 529–537.
5. Kolpakov F.A., Ananko E.A. (1999) Interactive data input into the GeneNet database. Bioinformatics. 15, 713–714.
6. Panich I., Makeeva E., Dromashko S. (1998) CHEDIBASE – interdisciplinary searching and base for the Chernobyl information. EUROMAB VI: Proc. MAB Sci. Symp. "Use and Conservation of Biological Resources", Minsk, Belarus, 16–20 September 1998. 194–203.

ON SPECIALIZATION "BIOINFORMATICS" IN THE NOVOSIBIRSK STATE UNIVERSITY AND HIGH COLLEGE OF INFORMATICS OF NOVOSIBIRSK STATE UNIVERSITY

[1] Kolchanov N.A., [2] Valishev A.I., [1] Popova N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: kol@bionet.nsc.ru
[2] High College of Informatics NSU
*Corresponding author

Key words: *bioinformatics, tertiary education, vocational education training, labour market analysis*

Introduction

Rapid development of informational biology needs to scale-up the training of specialists with both high and vocational education. For meeting this challenge, the Novosibirsk State University together with the High College of Informatics NSU have extended the training of the specialists, middle managers, in bioinformatics in order to supply the studies in the fields of molecular genetics, biomedicine, ecology, and selection in institutions with relevant specializations.

On the basis of Department of Natural Sciences NSU, we have organized a Chair "Information Biology" for training well-qualified specialists with the fundamental education in biology, mathematics, informatics, physics. NSU has a large experience in training the specialists - mathematical biologists, which will be a substruction of training at speciality «Information Biology». The specialists of this profile are callable in a wide variety of academic research organizations in Russia and abroad.

Modern biology became a producer of unprecedented bulk of experimental data that could not be digested without involvement of modern informational technologies and effective mathematical methods of data analysis and modeling of biological systems and processes.

In response to the necessity in large-scale training of specialists in the fields of informational biology, it is planned to organize a Chair of Information Biology (CIB) on the basis of Department of Natural Sciences of the Novosibirsk State University. The rough estimate of the number of students of the Faculty of Biology of Department of Natural Sciences, which will be enlisted annually at the Chair of Information Biology, should vary from at least 15 to at most 20-25 persons. The Chair will produce the graduates of the Department of Natural Sciences (DNS) of the Novosibirsk State University, speciality «Biology», with specialization in Information Biology. It is planned that the Chair will prepare young specialists for working at research scientific organizations dealing with fundamental biological problems, ecology, medicine, conservation of environment, etc., as well as studying the cross-disciplinary problems between biology and mathematics, computer science, physics, and chemistry.

For working in the field of Information Biology, the fundamental knowledge is necessary in such areas as biology, mathematics, informatics, and physics. With this respect, we suggest to train the specialists in information biology on the basis of integrative approach by using the educational resources of the Department of Natural Sciences and those of other departments of NSU (e.g., Department of Mathematics and Mechanics, Department of Physics, Department of Information Technologies). In the process of education, the Chair of Information Biology of Department of Natural Sciences NSU will serve as an integrating center.

The Chair of Information Biology will also provide organization of the educational process of the students and magisters of the other departments of NSU (Department of Mathematics and Mechanics, Department of Physics, Department of Information Technologies) who wish to specialize in information biology (by means of delivering additional or alternative disciplines, seminars, practicals, supervising diploma projects in cooperation with educating chairs of departments listed above). In the course of this education, the students and magisters of these departments will obtain additional knowledge and practical skills necessary for working in the field of information biology.

It is supposed that education of biologists who will further specialize at CIB, during the first three years will be based mainly on the standard programs adopted at the DNS of NSU. However, the modifications of the program should be done. Additionally, some courses should be changed for the alternative ones, which will be delivered by the staff of CIB. For the purposeful training of the graduates who intend to work in the definite area of information biology and who prepares the diploma project in this field, it is possible to use individual tutorial plans. At the 4-th and 5-th courses, the education will include delivering of the basic courses, special courses, seminars, practicals, etc. (with the total volume from 250 to 400 hours per year), preliminary diploma practice and diploma project.

The basic institute for the Chair of Information Biology will be the Institute of Cytology and Genetics SB RAS, where currently about 50 researchers work in this speciality. The institute has about 10 laboratories, which could support performance of diploma projects in this speciality. Diploma practice may be also performed at the other institutes of SB

RAS: Institute of Bioorganic Chemistry, Central Botanic Garden, Institute of Systematics and Ecology of Animals, Institute of Soil Science and Agrochemistry, Institute of Chemical Kinetics and Combustion, Institute of Thermophysics, Institute of Computational Technologies, Institute of Computational Mathematics and Mathematical Geophysics, Institute of Mathematics.

We suppose that specialization of the students – mathematical biologists at the Chair of Cytology and Genetics of DNS of NSU will be also conserved. However, the students – mathematical biologists will have an opportunity to get more qualified education in different subjects of biology, mathematics, bioinformatics and physics, which will be delivered to the students of the Chair of Information Biology.

At the Chair of Information Biology, we plan to organize the work on designing informational and software resources realized at the market of information technologies. Thus, this will enable the students to gain additional earnings and to achieve higher level of material support in comparison to the stipends delivered by NSU.

In the High College of Informatics NSU (HCI NSU), the novel specialization, "Bioinformatics", of the specialized secondary education is already available within the frames of the basic specialization № 2203, "Program software support of computing technique and automatic processes". The specialist working in the field of bioinformatics needs to have deep knowledge in biology, mathematics, informatics, and physics. To this aim, we plan to train specialists in bioinformatics on the basis of integrative approach by using educational resources both of NSU and the basic scientific research institution, the Institute of Cytology and Genetics of SB RAS.

For specialization «Bioinformatics», we have worked out *de novo* the programs of four novel courses: course of biology entering the block of general disciplines, course "Automated practice position" entering the block of special disciplines, the course "Computer modeling of biological processes", the course "Processing of experimental data and computer-assisted modeling in natural sciences". The authors of the courses mentioned above are the staff researchers of the Institute of Cytology and Genetics, of other institutions of SB RAS, and staff educators of the HCI NSU. For the course of biology, the researchers of the IC&G SB RAS A.A.Yushkova and N.A.Popova have prepared in the round methodic and tutorial books. All the courses are fitted up with methodological books and manuals, which are supplied by the packages of self-activity task books.

The special courses during specialization could be voluntary chosen. So, the narrow specialization could be achieved and realized in diploma project and in future working activities. The specialized courses "Computer methods of modeling and analysis of biopolymers", "International genetic databases", "Molecular-genetic systems and processes", "Bibliographic databases in biology", and "Modern methods in bioinformatics" are devoted to studying contemporary information technologies applied for analysis of molecular-genetic data in molecular genetics, molecular biology, and other directions of the modern biology. The content of courses is based on the analysis of contemporary scientific researches. It is illustrated by demonstration of specialized computer systems, by studying of methods and algorithms, which are necessary for experimental data treatment.

Among the characteristics of education are the usage of technology of the problem-oriented approach, which is based on the system of the learning-professional projects: the initiatory and basic projects at the 1st and 2nd courses, the project on specialization at the 3rd and 4th courses, which are performed by the students during the periods of summer field trip, diploma project at the 4th course. Thus, the student is constantly immersed into professional activities and monthly, he (or she) fulfils the individual task, the primary goal of which is to design the program product. Next, the transition is made from the educational goals to execution of the real projects and personnel professional products. To the important tasks of the projective form of education, we refer the attainment of writing the reports and documenting the software products, the abilities for working in a collective (first, in the educational one, then, in research or industrial ones), the ability to make a report about the tasks fulfilled and to advocate the personnel work against the Examining Board.

The field professional study and preparing of diploma projects of the students learning at the specialization «Bioinformatics», takes place at the Institute of Cytology and Genetics SB RAS and other basic organizations.

Analysis of the labor market by distributing questionnaires among the specialists of some organizations has proved that the specialist-bioinformatician is recallable by the organizations of research, technological, industrial, and other profiles of different patterns of ownership in the field of bioinformatics.

EDUCATIONAL COMPUTER PROGRAMS "MENDEL'S LAWS" AND "EXPERIMENTS WITH *DROSOPHILA MELANOGASTER*"

Berlizev A.A., Krasovitskiy A.M., Myasnikoff N.N., Biaysheva Z.M.

Institute of Mechanics and Mathematics and Department of Applied Mathematics,
al-Faraby Kazakh National University, Almaty, 480078, Republic of Kazakhstan, e-mail: aleks_kras@pisem.net

Key words: *education, computer program, Mendel's laws, Morgan's experiments*

Resume

Motivation: Educational programs in the form of virtual laboratories are most appropriate for lecture courses and individual study of the classical genetic laws.

Results: We produced the virtual laboratories "*Mendel's laws*" and "*Experiments with Drosophila melanogaster*". The "*Mendel's laws*" program allows one to cross plants of pea in the mono- and dihybrid modes. The program provides the pea forms and colors for the study. Statistical analysis of the results of these processes is also realized. The "*Experiments with Drosophila melanogaster*" program has four modes: "*Research of Genotype*", "*Cross-breeding*", "*Crossing-over*", and "*Gene engineering*". The program provides the *aristaless*, *straw*, *vestigial*, *ebony*, *curled wings*, *eyeless*, and *white* fruit fly alleles. Everyone can repeat these classical genetic experiments working in our virtual laboratories. The manuals accompany these labs.

Availability: http://www.dgstudios.com/nick

Introduction

Educational computer programs teach genetics basics and provide the student with many genetics experiments (see for example www.biodisc.com\genetics.htm). With this excellent genetics introduction, a series of experiments in a virtual laboratory may be executed quickly and easily. The educational programs of such type attract the attention of student by full imitation of real objects. Each of these programs is unique in the sense of functional logic of interface and educational provision kit. We have also elaborated our computer versions of the classical Mendel and Morgan experiments (Mendel, 1866; Morgan, 1927, 1932).

Methods and Algorithms

Both programs had been written in object-oriented language *Delphi*. We attended to the specific hierarchy of the program objects and their encapsulations. The logic algorithm of the programs introduce some "intelligence" into interface behavior. The computer resources (memory, CPU time, and system handles) have been adapted to these goals. The animation of fruit flies was reduced to *idle*-thread in order to force up efficiency of virtual experiments. The program produces for experiments a plenty of various hybrids of *D. melanogaster* by manipulation of both phenotype layers and genotype data. The *CoolAnimate* program component was created as a byproduct. It was used as a convenient intermediate between the graphical objects and the language of the script.

Implementation and Results

The "*Mendel's laws*" program provides the pea forms (round and wrinkled) and the pea colors (green and yellow) for the mono- and dihybrid virtual crossing. The statistical mean and dispersion are registered. The "*Experiments with Drosophila melanogaster*" program has four modes: "*Genotype Research*", "*Cross-breeding*", "*Crossing-over*", "*Gene engineering*". The program allows one to experiment with obviously expressed mutations such as *aristaless*, *straw*, *vestigial*, *ebony*, *curled wings*, *eyeless*, and *white*. At the very beginning it is quite enough for good representation of fundamental genetic laws. Sex chromosomes and autosomes are inherited by chance. The "*Crossing-over*" mode has an additional tool that exchanges some parts of homologous chromosomes. The "*Crossing-over*" mode allows student to realize virtual genetic mapping. The "*Research of Genotype*" mode shows graphically the interconnection between the fruit fly phenotype and genotype. The student can complete fruit fly's chromosomes within gamete or zygote nuclei and start the ontogenesis process in the "*Gene engineering*" mode. The historical row of genetic discoveries and classical experiments are saved in the similar design of the interface of both virtual laboratories.

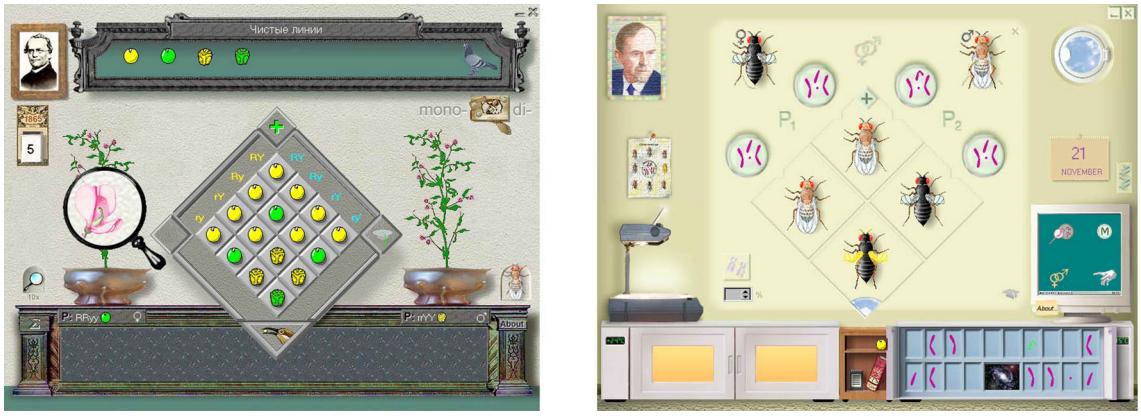


Fig. Interfaces of "Mendel's laws" and "Experiments with *Drosophila melanogaster*" in the crossing mode.

References

1. Mendel G. (1866) Versuche űber Pflanzenhybriden. Verhandl. naturforsch. Vereines Brűnn, **4**, Abhandlungen, 3-47.
2. Morgan T.H. (1927) *Experimental Embryology*, Columbia University Press, New York.
3. Morgan T.H. (1932) *The Scientific Basis of Evolution.* W.W. Norton and Company, Inc., New York.

MEDIATION OF HETEROGENEOUS INFORMATION RESOURCES IN THE GENE EXPRESSION REGULATION DOMAIN

[1] Kalinichenko L.A., [1] Briukhov D.O., [1] Zakharov V.N. , [2,3] Podkolodny N.L.

[1] Institute for Problems of Informatics, RAS, Moscow, Russia, e-mail: {leonidk,brd}@synth.ipi.ac.ru
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia, e-mail: pnl@bionet.nsc.ru

Key words: *subject mediator, heterogeneous information sources, semantic information integration, ontology, gene expression regulation*

Resume

Motivation: Semantic integration of heterogeneous information and procedural sources in bioinformatics is a necessary pre-requisite for efficient research.

Results: Application of the subject mediation approach in bioinformatics is shown for the gene expression regulation domain.

Availability: The results are available on request from the authors.

Introduction

A discriminative feature of molecular-genetic systems is their complex hierarchical and/or network organization. For instance, an organ consists of tissues, a tissue – of various cell types, a cell – out of compartments (i.e., cytoplasm, nucleus, vacuoles, etc.) that contain the macromolecules of DNA, RNA, and proteins. These macromolecules intensively interact with each other (they organize complexes, act in various reactions, move through cell compartments, cells, tissues, and organs, etc.), thus forming a composite net of interactions, namely, the gene network.

While solving concrete problems that are important in practice it is necessary to use a large number of heterogeneous, weakly structured molecular-genetical databases accumulating the results of numerous, complementary, intersecting, and probably contradictory experimental data. Databases on molecular-genetic information store the sequences, structures, 3D descriptions, attributive information, along with program software tools for data analysis, search of regularities, and prediction of different properties of objects, data reorganization, visualization, etc.

For efficient organization of research in the domain of bioinformatics it is required to organize properly the relevant information in specific research areas. One of the important outcomes of such organization would be provision of access to and querying of a large number of distributed information sources including various data on the primary and spatial structure of DNA and RNA macromolecules, proteins and their complexes as well as data on peculiarities of their interactions with each other.

Such data usually are semistructured. For their processing, a significant amount of additional metainformation, complex semantic analysis combining various methods may be required. The problem becomes even more complicated because data stored in different sources are obtained for different research entities, with different precision describing real processes in the living organism.

To provide for semantic integration of nonsystematic population of autonomous information sources kept by different information providers into a well-structured information collection it is required to create the global unified representation of the existing information sources and services. To reach that it is proposed to form a special middleware consisting of the *subject mediators*. For each subject mediator, the application domain model is to be defined by the experts in the field. This model may include specifications of data structures, terminologies (thesauri), concepts (ontologies), methods applicable to data, processes (workflows), characteristic for the domain. These definitions constitute specification of a subject mediator. Due to that, mediators provide a uniform query interface to the multiple data and procedure service sources, thereby freeing the users from having to locate the relevant sources, query each one in isolation, and combine manually the information from them.

We develop a mediator for integration of heterogeneous molecular-genetic data in the area of gene expression regulation. The three level mediator architecture consists of federated, local and intermediate layers. The federated layer keeps subject mediator specifications, such as ontological definitions of the subject domain, schema description defining structural (types, classes, attributes) and functional (e.g., facilities for semantic data analysis and predictions, knowledge discovery based on the automatic methods) capabilities of the mediator. The local layer represents canonical specifications of the heterogeneous sources registered at the mediator. The intermediate layer defines a mapping of the source specifications into the specifications of the mediator.

Advantages of the proposed approach include the following:

Semantic integration of heterogeneous information collections can be reached by taking into account structural, value, semantic, quality data heterogeneity;

Users should know only subject definitions that contain concepts, structures and methods as defined by the community.

Querying the subject definitions, users have integrated access to all information registered at the mediators up to the moment of a query.

Personalization providing convenient views for specific groups of users can be formed above the subject definitions. This process is independent of the existing collection and their registration.

The mediator structure includes the metainformation base, tools for information sources registration, query interpretation facilities, facilities for collecting the query results and providing them to users.

The mediator for gene expression regulation

The model of the subject domain (gene expression regulation) has been developed. The model has a multilevel structure and includes ontological definition of the related concepts and thesauri, definition of information structuring, types of experiments, data analysis methods, as well as the related models of the respective theory.

The mediator is oriented on a broad class of problems. The intuition behind them can be provided by an example sequence of interrelated queries to the mediator that are intended for preparation of the training samples of regulatory regions, which may be used by recognition programs: to output the set of transcription factor binding sites sequences, which have a definite type of DNA-binding domain, search for transcription factors corresponding to the proteins found, search for transcription factor binding sites; search for the sequences of pre-ordered length including relevant transcription factor binding sites.

Examples of the ontological definitions represented in the metainformation base:

Name	"protein"
Definition	"A large molecule composed of one or more chains of amino acids in a specific order; the order is determined by the base sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, function, and regulation of the bodys cells, tissues, and organs, and each protein has unique functions. Examples are hormones, enzymes, and antibodies."
Name	"transcription factor"
Definition	"A protein that regulates transcription after nuclear translocation by specific binding with DNA or by stoichiometric interaction with a protein that can be assembled into a sequence-specific DNA-protein complex."
Part-of	"transcription complex"
Subclass-of	"regulatory protein"
Subclass-of	"protein"

Fig. 1 shows the fragment of the mediator schema specification in UML notation. These specifications are used to illustrate heterogeneous sources registration and querying of the mediator.

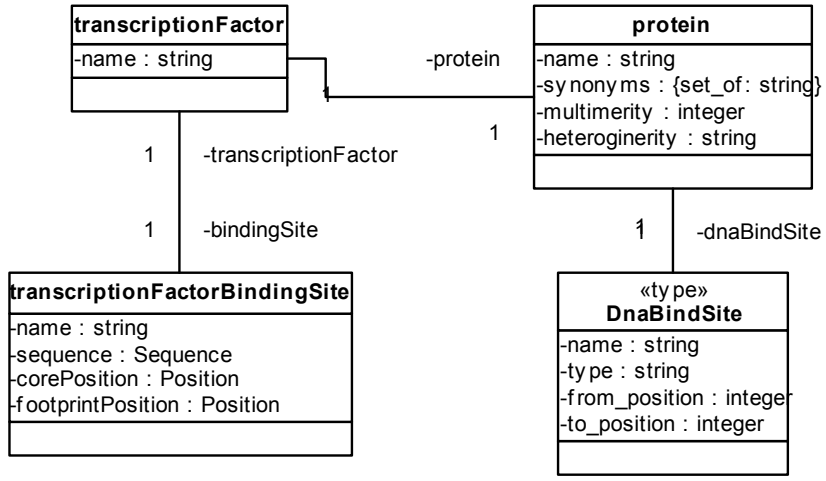


Fig. 1 The fragment of mediator schema specification in the UML notation.

Information Sources

Initial set of information sources to be registered at the mediator includes:

The database TRRD developed at the Institute of Cytology and Genetics, unique informational resource that has neither world-wide analogs and that contains information about structural and functional organization of extended transcription regulating regions of eukaryotic genes and their expression. A subset of the TRRD schema using in this paper contains classes *sites, factors* and types *SITES, FACTORS*. (Kolchanov, 2002a)

The database SWISSPROT contains an information about the structure and functions of genes, about their domain structure, sequences, etc. A subset of the SWISSPROT schema using in this paper contains class *sprotein* and types *SProtein, Description, Feature, Dna_bind*.

The databases EMBL/GenBank accumulate information about the sequences DNA, RNA, their exon-intron structure, and other functional layout.

The database Medline/PubMed stores bibliography that is necessary for supporting and verifying the data presented.

Registration of the information sources in the mediator

The process for registration of heterogeneous information sources at the subject mediator is based on the LAV (Local as view) approach. In LAV the registered collections schemas are considered as materialized views above virtual classes of mediator. In the specific registration method developed (Briukhov et al., 2001) the materialized views mentioned are designed applying compositional development method (source schema is treated as a specification of requirements and class schemas of the mediator are treated as component specifications). This approach is intended to cope with a dynamic, possibly incomplete set of sources. Sources may change their exported schemas, become unavailable from time to time. To disseminate the information sources, their providers should register them at a respective subject mediator. Such registration can be done concurrently and at any time. Specific methods and tools supporting process of information sources registration have been developed to make mediators scalable with respect to a number of sources involved.

During the registration a local source class is modeled as a set of instances (objects) of the class instance type, and the description of the source in terms of the mediator schema specifies the constraints on the class instances to be admissible for the subject mediator. The process of registration includes the ontological-based reconciliation of the application contexts of the registered sources and that of the mediator, identification of relevant classes of the mediator schema, constructing of the most common reducts for the mediator type specifications and respective types of the sources, constructing of views, specifying source class constraints in terms of the mediator classes.

Due to the space limit we cannot show here how contexts are reconciliated and types of the mediator are represented by the related types of the sources by means of the so-called concretizing reducts (Kalinichenko et al., 2000). We only show how views are expressed by means of the inverse rules (Duschka, Genesereth, 1997) expressing classes of the mediator through classes of sources (first rule relates protein to a class in SWISSPROT, two other rules relate *transcriptionFactor* and *transcriptionFactorBindingSite* to the respective classes in TRRD):

protein(p/Protein_SProtein) :- sprotein(p/Protein_SProtein)

transcriptionFactor(t/TranscriptionFactor_FACTORS) :- factors(t/TranscriptionFactor_FACTORS)

transcriptionFactorBindingSite(s/TranscriptionFactorBindingSite_SITES) :- sites(s/TranscriptionFactorBindingSite_SITES)

Query rewriting in terms of the sources

We consider here an example of a query to the mediator:

Display the transcription factor binding sites with the definite types of DNA binding domain

In the mediator's canonical model this query is expressed as:

transcriptionFactorBindingSite(s) & transcriptionFactor(t) & protein(p) & s.transcriptionFactor = t & t.protein = p & p.structure.type = "HOMEBOX"

After query rewriting applying the inverse rules above, we get the query:

sites(s/TranscriptionFactorBindingSite_SITES) & factors(t/TranscriptionFactor_FACTORS) & sprotein(p/Protein_SProtein) & s.transcriptionFactor = t & t.protein = p & p.structure.type = "HOMEBOX"

This query is implemented by a subquery SQ1 to TRRD and a subquery SQ2 to SWISSPROT with the remaining postprocessing in the mediator SQ3:

SQ1(s,t):- FACTORS(t/TranscriptionFactor_FACTORS) & SITES(s/TranscriptionFactorBindingSite_SITES) & s.transcriptionFactor = t

SQ2(p):- sprotein(p/Protein_SProtein) & p.structure.type = "HOMEBOX"

SQ3(s,t,p) :- SQ1(s,t) & SQ2(p) & t.protein = p

Conclusion

The paper is of the "work in progress" kind. It shows how the subject mediation approach can be applied in bioinformatics. Gene expression regulation domain has been chosen to define an example of the subject mediator. The paper briefly introduces a notion of a subject mediator and explains how it can be defined. Issues of heterogeneous sources registration at

the mediator and query rewriting in terms of registered sources are given in more details. The paper shows benefits that can be obtained applying the subject mediation approach.

An approach developed will be used as the tool for integration of information-software resources entering the integrated system on gene expression regulation, GeneExpress, which is being developed at the Institute of Cytology and Genetics of SB RAS (Kolchanov, 2002b). This system integrates heterogeneous program-informational resources (a large bulk of databases, and hundreds of programs). We believe that developed by us technology of mediators is an adequate tool to accomplish the task that faces us.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Project № 65).

References

1. 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. Nucl. Acids Res. 2002a, 30, 312-317.
2. Kolchanov N.A., Podkolodny N.L., Ananko E.A. etc. Integrated system on gene expression regulation GeneExpress – 2002. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002). 2002b.
3. Briukhov D.O., Kalinichenko L.A., Skvortsov N.A. Information sources registration at a subject mediator as compositional development Proceedings of the Fifth East European Symposium on Advances in Databases and Information Systems (ADBIS'01), Springer, LNCS, 2001.
4. Duschka O., Genesereth M. Answering Queries Using Recursive Views. In Principles Of Database Systems (PODS), 1997.
5. Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N. Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections. Proc. of the Second Russian National Conf. on "Digital Libraries: Advanced Methods and Technologies, Digital Collections, Sep. 26-28, 2000, Protvino.

TOWARDS A METRICAL SPACE OF BIOLOGICAL SEQUENCES

[1,2] Heymann S., [2] Gabrielyan O.R., [3] Ghazaryan G.G., [3] Danielyan E.A., [3] Hakobyan G.G., [3] Hakobyan G.O.

[1] Humboldt-University Berlin, Institute of Computer Science, Berlin, Germany
[2] Kelman GmbH, Berlin, Germany
[3] Yerevan State University, Yerevan, Armenia
*e-mail: heymann@dbis.informatik.hu-berlin.de

Key words: *amino acid sequence, metrical space, constructed "continuous" distance*

Resume

Motivation: Character strings representing the monomer succession in linear polymers are being scrutinized from many different semantic aspects. However, as any abstraction, sequence perception preferentially by means of discrete residue algorithms necessarily remains one-sided. It doesn't facilitate the elucidation of those biologically meaningful molecular traits that depend, generally speaking, on a biopolymer's chain continuity and integrity. To widen the scope of bioinformatics towards a desirably more holistic compilation of both the intrinsic molecular properties and some superordinate correlations in the living world, it is legitimate to ask in how far the use of *functions in the mathematical sense of the word* may assist researchers with addressing sequence-depending problems better than discrete maths does to date.

Results: In this paper, we proposed a method for mapping the set of biological sequences as words with characters from a given alphabet into the metric spaces, and for using the properties of this metrics of these metrical spaces for getting the distance between the words of the set.

Availability: The algorithms are described in the paper. As an application, we obtained the "continuous'' distance between some natural amino acid sequences (academic and commercial).

Introduction

The distance between two sequences $a = a_1a_2...a_n$ and $b = b_1b_2...b_m$ is classically defined by the minimum sum of distances of the characters $d(a_i,b_j)$, $d(a_i,\varnothing)$, and $d(\varnothing,b_j)$, where a_i, b_j are some characters, and $\varnothing$ is the "empty" character $(1 \le i \le n, 1 \le j \le m)$, where the minimum is extended over all alignments of a with b. In this definition, characters are taken as isolated elements in a sequences and no neighborhood influence on a_i, b_j or $\varnothing$ is being considered. In contrast to that, if a and b were brought in accordance to functions $a(t)$ and $b(t)$ of a continuous argument t, one observes for fixed argument values t_0 a "distance" between $a(t_0)$ and $b(t_0)$ that is influenced by function values around t_0.

In monograph by S.M.Ulam [1], the evolutionary distance of any two words was defined as the minimal sum of operation costs (for insertion, deletion, mutation, respectively) necessary for transforming word a stepwise into word b (commonly known as *minimum edit distance*).

In 1970, Needleman and Wunsch [2] introduced a dynamic programming algorithm for the global similarity assessment/alignment of two words with a constant gap penalty $g(k) = \alpha$ for all positions k opposite to an $\varnothing$. Importantly, insertions and deletions are not allowed to be adjacent. Dynamic programming was further developed by Sankoff [3] and independently applied by Sellers [4] and Wagner and Fisher [5], to determine the distance of real amino acid successions. This procedure requires *mxn* operation steps for word lengths of *m* and *n*, respectively.

With the appearance of the Smith–Waterman algorithm, FASTA, BLAST, and HMM Search, local alignment detection methods and tools paved the way to mass sequence fund scrutinize, together with heuristic and probabilistic concepts, models and (pre-) processing steps. A selection of bioinformatics textbooks reflects this success story with all the *pros* and *contras* these algorithm families revealed in practice. Reliability and precision questions are instantly subject to open-end investigations, to approach the goal of *near-optimal alignments*.

The history of functional ("continuous") distances of two numerical successions, the elements of which are "letters" in a numerical alphabet V with a given distance matrix D, started evidently from papers by Kantorovich and Rubinstein [6] and Wasserstein [7]. In the book [8], papers are subsummarized that were dedicated to the method of *Time Warping*. The common nucleus of these studies is a calculus of the "continuous" distance between two curves in different phase spaces: curve distances are the resulting values that arise for all possible transformation paths of the curves in another by the aid of a given distance function $G(x,y)$ and of given numeric equivalents (costs) for all transformation operations.

We haven't found any hint for Time Warping approaches to calculate amino acid sequence distances in literature, and this seems quite plausible. In the common perception, polypeptides are depicted sufficiently well as amino acid sequences, i.e. are regarded plainly as a matter of discrete nature. As outlined above, all the alignment and pattern search techniques available today were stimulated just on the background of this paragonic understanding. Therefore, the search for alternatives was apparently superfluous. Apart from polemics: why should one direct efforts onto transforming character strings first into continuous functions and then rediscretize those for Time Warping-based distance computations of isolated curve dots? It is clearly worth the attempt to dediscretize character strings, by an injective projection into a metrical space of functions. Prior to constructing the according functions, the best relocation (or one of the best relocations) of $a = a_1a_2...a_n$ and $b = b_1b_2...b_m$ is to be found by shifting the sequences as wholes relative to each other and by adding "empty" characters in a way that the word pair $(\tilde{a},\tilde{b})$ derived from (a,b) were of minimum distance and equal length. In this process the longest identical word segments in a and b will be superpositioned and fixed as well.

Methods and Algorithms

In this paper, we have used the methods of mathematical analysis, and numerical methods with some modification. The algorithms outlined in this work comprise a consistent description formalism of linear biopolymers in terms of continuous functions. The proposed algorithmic components represent carefully designed methodical prerequisites for a number of applications. The perhaps most advantageous feature of the new formalism consists in the considering of neighborhood impact on the trajectory through each single node. This way, even "invisible" local messages imprimed during evolution onto a given chain are taken over, thus modulating the course of the function. The latter effect remains in power for implicit information that cannot be dissected and understood in terms of similarity character-by-character or by statistical methods, but affects the metrical behavior. The before mentioned discriminatory power between naturally occurring and artificial strings (the latter within the variability of the main statistical attributes) gives a premonition of what is still fully hidden behind a protein's primary structure.

Implementation and Results

An algorithm has been constructed for the transition from a character alphabet U to a numerical alphabet V. This way, a certain numerical sequence is being assigned to each amino acid sequence. At the next procedural step, the numerical sequence will get dediscretized. As the result, any character string $a = a_1a_2...a_n$ will be replaced with a function $a(t)$. At the subsequent step the distance matrix will be dediscretized. In the result, any matrix M is being assigned to a function $G(x,y)$ of two independent variables. Next, we deduce the construction algorithm for a pair of functions $\{\varphi(t),\psi(t)\}$ necessary for compression and expansion of the according intervals of the functions $a(t)$, $b(t)$ to compare and for the penalty function $\lambda_{\varphi\psi}(t)$. Finally, by the aid of $G(x,y)$, $\varphi(t),\psi(t)$ and $\lambda_{\varphi\psi}(t)$, the distance between $a = a_1a_2...a_n$ and $b = b_1b_2...b_m$ is being determined according to

$$d(a,b) = c(m,n,s)\cdot \inf_{\varphi,\psi} \int_0^T G\left[a(\varphi(t)),b(\psi(t))\right]\lambda_{\varphi\psi}(t)dt ,$$

where the lower boundary is chosen from all the allowed $\varphi, \psi, \lambda_{\varphi\psi}$; $c(m,n)=\frac{m+n-2s}{m+n}$, s is the length of the longest coinciding segment in a and b. The inner functional terms $\{\varphi(t), \psi(t), \lambda_{\varphi\psi}(t)\}$ will be constructed according to an algorithm that ensures a value of the integral in the right-hand part of the above relation "near" to the minimum value.

Discussion

The numerous possibilities to modulate the course of a function, e.g. by dictating the values of the 1st, 2nd, and higher-order derivatives in dependence on peculiarities like post-translational modifications a node may have, deserve (and require) subtle investigations. In the perspective, evolutionary, physicochemical and other biologically meaningful protein properties will be adequately projected into metrical spaces of continuous functions. With this respect, the chances of knowledge gain are expected to be superior as compared to discrete methodologies.

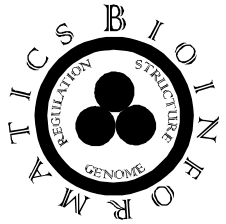
Acknowledgements

The authors are grateful to Kelman GmbH for the setting of the problem and for financial support. The authors are also thankful to referees for useful comments and criticism.

References

1. Ulam S.M. (1972). Some Combinatorial Problems Studied Experimentally on Computing Machines. S.K.Zaremba, New York: Academic Press.

2. Needleman S.B., Wunsch C.D. (1970). A general method applicable to the search for similarities in the amino acid sequences of two proteins. J. Mol. Biol. 48, 443-453.
3. Sankoff D. (1972). Matching sequences under deletion – insertion constraints. Proc. Natl Acad. Sci. USA. 68, 4-6.
4. Sellers P.H. (1974). An algorithm for the distance between two finite sequences. J. Combinator Theor. A16, 253-258.
5. Wagner R.A., Fischer M.J. (1974). The string-to-string correction problem. J. Assoc. Comput. Mach. 21, 168-173.
6. Kantorovich L.V., Rubinstein G.S. (1957). On a function space and certain extremum problem. Dokl. Akad. Nauk SSSR. 115(5), 1058-1061.
7. Wasserstein L.N. (1969). Markov processes over denumerable products of spaces describing large systems of automata. Problems Information Transmission. 5, 47-52.
8. Sankoff D., Kruskal J.B.(1999). Time Warps, String Edits and Macromolecules. CSLI Publications, ISBN 1-57586-217-4 (originally published 1983 by Addison-Wisley, Reading, MAS).
9. Baxevanis A.D., B.F.F. (2001). Oullette Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Second Edition. Wiley-Liss, ISBN 0471383910.
10. Mount D.W. (2001). Bioinformatics: Sequence and Genome Analysis. Gold Spring Harbor Laboratory, ISBN 0879695978.
11. Smith T.F., Waterman M.S., Fitch W.M. (1981). Comparative biosequence metrics. J. Mol. Evol. 18(1):38-46.
12. Waterman M.S. (1983). Sequence alignment in the neighborhood of the optimum with general applications to dynamic programming. Proc. Natl Acad. Sci. USA. 80, 3123-3129.
13. Vingron M., Waterman M.S. (1994). Sequence alignment and penalty choice: Review of concepts, case studies and implications. J. Mol. Biol. 235, 1-12.
14. Pevzner P.A. (2000). Computational Molecular Biology: An Algorithmic Approach. MIT Press, ISBN 0262161974.
15. Baxevanis A.D., B.F.F. (2001). Oullette Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Second Edition. Wiley-Liss, ISBN 0471383910.
16. Gibas C., Jambeck P. (2001). Developing Bioinformatics Computer Skills. O'Reilly and Associates, ISBN 1565926641.



BGRS 2002

OTHER TOPICS RELATED TO BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE

GENEEXPRESS-2002: AN INTEGRATED SYSTEM ON GENE EXPRESSION REGULATION

Kolchanov N.A. [1]*, ***Podkolodny N.L.*** [1,2], ***Ananko E.A.*** [1], ***Ignatieva E.V.*** [1], ***Podkolodnaya O.A.*** [1], ***Stepanenko I.L.*** [1], ***Merkulova T.I.*** [1], ***Lavryushev S.V.*** [1], ***Grigorovich D.A.*** [1], ***Kochetov A.V.*** [1], ***Orlova G.V.*** [1], ***Titov I.I.*** [1], ***Vishnevsky O.V.*** [1], ***Orlov Yu.L.*** [1], ***Ivanisenko V.A.*** [1], ***Vorobiev D.G.*** [1], ***Oshchepkov D.Yu.*** [1], ***Omelyanchuk N.A.*** [1], ***Pozdnyakov M.A.*** [1], ***Afonnikov D.A.*** [1], ***Matushkin Yu.G.*** [1], ***Likhoshvai V.A.*** [1], ***Ratushny A.V.*** [1], ***Katokhin A.V.*** [1], ***Turnaev I.I.*** [1], ***Proskura A.L.*** [1], ***Suslov V.V.*** [1], ***Nedosekina E.A.*** [1]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: kol@bionet.nsc.ru
[2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
*Corresponding author

Key words: *gene, genome, gene networks, regulation, expression, transcription, splicing, translation, DNA, RNA, proteins, regulatory sequences, databases, integration, knowledge discovery*

Resume

Motivation: A rapid discovering of experimental data on regulation of gene expression and their accumulation in various databases accompanied by development of numerous and diverse software tools for their analyses demands integration of the available informational and software resources.

Results: An Internet-accessible system GeneExpress-2.1 aimed for integration of informational and software resources of regulation of gene expression has been designed and developed.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/.

Introduction

Currently, the number of databases on gene expression and a variety of software for the analysis of these data are growing fast. An Internet-accessible system GeneExpress-2.1 is being developed for accumulation of experimental data, their analysis, and navigation through integrated software and informational resources related to regulation of gene expression. It integrates a large amount of databases and hundreds of programs for processing the data on the structure–function organization of DNA, RNA, proteins, and gene networks together with other informational resources important for gene expression regulation. From its first version (Kolchanov at al., 1998a; 1998b), GeneExpress is intensively developing (Kolchanov at al., 1999; 2000). This paper briefs the state of GeneExpress –2.1 in 2002. The system is widely and actively used for computer analyses of various organizational levels of molecular genetic systems. Descriptions of its individual modules in more detail are available in other papers included in *Proceedings of BGRS'2002.*

The structure of the **GeneExpress-**2.1 corresponds to the natural hierarchical organization of molecular genetic systems, containing the following levels: (1) **DNA level,** (2) **RNA level,** (3) **protein level,** and (4) **gene network level.** Each module contains (1) experimental data represented as a database or a sample; (2) program for data analysis; (3) results of an automated data processing; and (4) tools for graphical representation of these data and results of the data analyses. For access to the databases of GeneExpress-2.1 system, RDBMS (Relation Database Management System) ORACLE 9i, and Sequence Retrieval System (SRS 6.0) are used.

1. Informational and software resources on DNA structure–function organization

1.1. Transcription Regulatory Regions Database (TRRD)

TRRD is designed for accumulation of experimental information on the structure–function organization of regulatory regions of eukaryotic genes. It is a unique informational resource on long gene transcription regulatory regions. In addition to description of regulatory region itself, it provides (a) description of **the hierarchy of all the regulatory units** included into a described regulatory region (such as transcription factor binding sites, promoters, enhancers, silencers, etc.); (b) information on **expression patterns** of the genes described; and (c) information on **physiological systems, organs**, and **cell types** wherein these genes are expressed. The new release of **TRRD-6.0** contains ***interferon-inducible genes; erythroid-specific genes; genes of lipid metabolism*** in liver, adipose tissue, at the cell and organismal levels (cholesterol regulation, and leptin hormone regulation, lipid exchange between lipoprotein blood particles); ***glucocorticoid-inducible gene;, cell cycle–dependent genes; genes of the endocrine system; heat shock-regulated genes;*** redox sensitive genes; iron metabolism genes; macrophage-expressed genes; apoptosis genes; ***and plant genes***. **TRRD-6.0** comprises descriptions of 1460 genes, 6867 sites, and 2243 regulatory regions. This database is supported constructed using **ORACLE 9i,** and Sequence Retrieval System (**SRS 6.0**) is used to access **TRRD-6.0.** A novel version of the **TRRD Viewer** (release 2.0) implemented as a Java-applet allows the regulatory gene regions described in TRRD to be visualized.

1.2. Programs for recognizing regulatory elements involved in controlling the transcription

GeneExpress-2.1 has a large set of original programs for recognition and various analyses of transcription factor binding sites and promoters as well as study of specific contextual and structural DNA features in gene regulatory regions, exemplified below.

Resource	Description
BinomSite	Searching for potential transcription factor binding sites (TFBS) using a binomial criterion for estimating similarity scores between regions of a sequence analyzed and the TFBS sequences described in TRRD **URL**: http://wwwmgs.bionet.nsc.ru/mgs/programs/mmsite/
MMSite	Simultaneous usage of the entire set of recognition methods for detecting TFBSs. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/programs/multalig/
ARGO_Viewer	Recognition of promoters of tissue-specific gene groups basing on the analysis of the presence of specific quasi-invariant oligonucleotide motifs detected using the program **ARGO.** **URL**: http://wwwmgs.bionet.nsc.ru/mgs/programs/argo/argo_viewer.html
RGSiteScan	Searching for TFBSs basing on the recognition group approach. **URL**: http://wwwmgs.bionet.nsc.ru/mgs/programs/yura/RecGropScanStart.html
KD_Prom	Recognition of RNA PolII promoters from their contextual patterns determined using knowledge discovery and data mining in TRRD. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/programs/recon2/
ProGA	Recognition of RNA PolII promoters. **URL**: http://wwwmgs.bionet.nsc.ru/mgs/programs/proga/
BLAST_Promoter	Recognition of the RNA PolII promoters basing on the **BLAST** search for homology with the promoters described in the TRRD. **URL**: http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/blast.html
Recon	Searching for potential nucleosome formation sites basing on the nonuniformity of dinucleotide context within local promoter regions. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/

2. Informational and software resources on RNA structure–function organization

To solve a variety of problems on analyzing RNA structure–function organization, a number of databases and software tools were developed with GeneExpress-2.1 and united in the module **RNA Integration Level.** It comprises (i) a number of programs for calculation of RNA secondary structure and evaluation of the secondary structure formation potential and (ii) the knowledge base on structure–function organization of leader mRNA sequences.

Resource	Description
Program **Garna**	Applying genetic algorithm to predict the secondary structures displaying seals energies and visualize them. URL: **http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna/**
Program **MatrixSS**	Calculation of E score, a contextual characteristics reflecting the potential for forming RNA secondary structure compared with random sequences. URL: **http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna/MatrixSS.html**
Knowledge base **LEADER_RNA**	LEADER_RNA is a tool to evaluate mRNA translational properties. Contains a database with samples of 5'UTR sequences of high- and low-expressed mRNAs of mammals, dicot, and monocot plants. These sequences are used as training samples for the computer system. This knowledge base contains also (1) description of the discovered mRNA properties that may be used to discriminate between the high- and low-expressed mRNAs and (2) programs predicting mRNA translational efficiencies from significant contextual and structural characteristics of mRNA 5'UTRs (C codes for prediction of mRNA translation level). URL: **http://wwwmgs.bionet.nsc.ru/mgs/gnw/leader/**

3. Informational and software resources on the structure–function organization of proteins

A number of databases and software tools, forming the modules **Protein Integration Level** of **GeneExpress-2.1,** have been developed for solving the problems related to analyses, structure, function, and evolution of proteins. This module contains the databases on (i) expanded annotation of the EnPDB-compiled structures, (ii) active sites of the PDBSite-compiled proteins, (iii) protein and peptide sequences obtained by artificial *in vitro* selection (**ASPD**) as well as the programs for (iv) searching the protein spatial structure for re regions similar to **PDBSiteScan**-compiled active centers and (ii) detecting and analyzing the coordinatively fixed amino acid substitutions (CRASP, Gene Network Level)

Resource	Description
Database **EnPDB**	Expanded options for indexed search for information in the PDB databank entries **URL:** http://wwwmgs.bionet.nsc.ru/mgs/gnw/enpdb/
Database **PDBSite**	Information on the spatial structure and physicochemical properties of 4723 active protein sites annotated in PDB. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/
Program **PDBSiteScan**	Searching for active sites in protein spatial structures according to the sequence and spatial arrangement of amino acid residues. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html
Database **ASPD**	Information on peptide and protein sequences produced an *in vitro* selection. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd/
Program **CRASP**	Detection of coordinated amino acid substitutions in protein families and analysis of their physicochemical

	characteristics. **URL:** http://wwwmgs.bionet.nsc.ru/mgs/gnw/crasp/

4. Informational and software resources on the structure–function organization and operation dynamics of gene networks

This module of GeneExpress-2.1, **Gene Network Level,** comprises the data on structure–function organization of gene networks, tools for their visualization, and software for simulation of gene network dynamics. It includes the following components.

Resource	Description
Database GeneNet	Information on the structures of gene networks, compartmentalization of their components at the levels of cells and overall organism, functional interactions of gene network components, signal transduction pathways, and metabolic pathways. URL: http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/GeneNet-0002.shtml
Program GeneNet Viewer	Visualization of gene networks as interactive graphical layouts; analysis of the structure–function organization of gene networks; and analysis of interspecies differenced in gene network organization. URL: http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/applet_genenet_viewer.shtml
Knowledge base GeneNet Model	Sets of differential equations describing dynamics of reactions, reaction rate constants, and initial concentrations of the components for dynamical models of three gene networks, namely, gene network regulating (1) lipid metabolism, (2) erythrocyte differentiation and maturation under the effect of EPO, and (3) activation of macrophages by IFN-γ and LPS. URL: http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/
Program GeneNet Modeling	Simulation of gene network dynamics; analysis of gene network operation when concentrations of individual components or rates of individual reactions are changed; simulation of effects of additional components on the gene network operation; and simulation of mutations. URL: http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/modelling.shtml

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65) US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049).

References

1. Kolchanov N.A., Ponomarenko M.P., Kel A.E., Kondrakhin Yu.V., Frolov A.S., Kolpakov F.A., Kel O.V., Ananko E.A., Ignatieva E.V., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Babenko V.N., Vorobiev D.G, Lavryushev S.V., Ponomarenko Yu.V., Kochetov A.V., Kolesov G.B., Podkolodny N.L., Milanesi L., Wingender E., Heinemeyer T., Solovyev V.V. (1998a). GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences of the eukaryotic genome. ISMB,. 6:95-104. MEDLINE PMID: 9783214; UI: 98456543.
2. Kolchanov N.A., Ponomarenko M.P., Kondrakhin Yu.V., Frolov A.S., Kolpakov F.A., Kel A.E., Kel-Margoulis O.V., Ananko E.A., Ignatieva E.V., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Babenko V.N., Vorobiev D.G., Lavryushev S.V., Grigorovich D.A., Ponomarenko J.V., Kochetov A.V., Kolesov G.B., Podkolodny N.L., Wingender E., Heinemeyer T., Milanesi L., Solovyev V.V., Overton O.K. (1998b). GeneExpress system: description, analysis, and recognition of regulatory sequences in eukaryotic genomes. Proc. I Intern. Conf. on Bioinformatics of Genome Regulation and Structure, BGRS'98, Novosibirsk–Altai Mountains, August 24-31, 1998, 71-76.
3. Afonnikov D.A. (2002). Contribution of coadaptive substitutions to the stability of physicochemical properties of ATP-binding sites in protein kinases. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
4. Afonnikov D.A., Oshchepkov D.Yu., Kolchanov N.A. (2001). Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with coordinated substitutions. Bioinformatics. 17, 1035-46.

ANALYSIS OF THE SECONDARY STRUCTURE AND NUCLEOSOMAL POTENTIAL OF *NOT* I SITES OF THE HUMAN GENOME

[1] Matushkin Yu.G., [1] Levitsky V.G., [1] Likhoshvai V.A., [1] Vishnevsky O.V., [2,3] Kutsenko A.S., [2,3] Protopopov A.I., [2,3] Zabarovsky E.R., [1] Kolchanov N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: mat@bionet.nsc.ru
[2] Center for Genomics and Bioinformatics, Karolinska Institute, 171 77, Stockholm, Sweden
[3] Microbiology and Tumor Biology Center, Karolinska Institute, 171 77, Stockholm, Sweden
[*]Corresponding author

Key words: *computer analysis, nucleosomal potential, local complementation index, oligonucleotide frequencies, human genome sequences, NotI sites*

Summary

Motivation. Complete sequencing of the human genome posed the task of investigation of the functional properties of genomic sequences. An important theoretical and practical problem is to deduce from the primary structure as many traits affecting the behavior of sequences in experimental studies as possible.

Results: We have found significant differences in the oligonucleotide compositions of two pairs of sequence samples from chromosome 3: clones positive or negative with respect to hybridization and clones positive or negative with respect to PCR. For all sequences, the nucleosomal potential (NP) has a global minimum near the *Not*I site. The NP of chromosome 21 clearly tends to decrease from the centromere to telomere. This trend is hardly noticeable in chromosome 22. The samples $Clon^+$ и $Clon^-$ from chromosome 21 differ significantly (significance exceeds 99.99%) in the local complementation index (LCI) in the range of 4000–4300 from the 5'-end or 5700–6000 downstream of the *Not*I site. The LCI profile has a maximum at the *Not*I site for all sequences examined.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/; (Nucleosomal potential) http://wwwmgs.bionet.nsc.ru/mgs/programs/argo (ARGO).

Introduction

A human genomic DNA library has been produced. It contains recognition sequences of the *Not*I restriction enzyme and adjoining flanks. A total of 22,551 unique *Not*I flanking sequences were generated, which cover 16.2 Mb of the human genome. These sequences exhibit differences with regard to hybridization, PCR, etc. The difference may be determined by the features of the sequences themselves, in particular, their ability to form secondary structures. Because of the association of *Not*I sites with genes, the study of putative regulatory properties of the neighborhood of these sites appears to be an urgent task.

Materials, Methods, and Algorithms

A human genomic DNA library was produced, which contains *Not*I restriction enzyme recognition sequences and adjoining flanks (Kutsenko *et al.,* 2002). Eight samples from chromosomes 3, 21, and 22 were investigated. Two samples of chromosome 3–associated *Not*I clones were constructed using PCR. All plasmids were subdivided into two samples, which showed or failed to show the specific PCR product and were designated as PCR^+ and PCP^-, respectively. Under the experimental conditions used, the efficiency of PCR did not depend on the insert length.

The second pair of samples was based on DNA–DNA hybridization. Plasmid DNAs of *Not*I clones were spotted onto glass slides at equimolar concentrations for a microassay. Hybridization probes were prepared according to the CODE protocol (Li et al., 2001). As the microassay often provides ambiguous results, the clones were subdivided into two samples according to the presence or absence of the specific signal. They are denoted as HYB^+ and HYB^-. Under the experimental conditions used, the efficiency of the microassay hybridization did not depend on the insert length or presence of Alu repeats. We investigated two more samples for chromosomes 21 and 22. They were differentiated by their presence in or absence from the *Not*I library. Those present were denoted as "$Clon^+$", or "alive", and those absent, as "$Clon^-$", or "dead". Actually, this trait describes the accessibility of the molecule for restriction endonucleases and/or the circularization ability of the region from the *Not*I site to the nearest *Bam*HI (1–15 kb) (Kashuba et al., 1997).

We used several algorithms and programs developed at the Institute of Cytology and Genetics, Novosibirsk: the package for calculating local complementation index, which is a measure of the number of high-energy secondary structures (Likhoshvai, Matushkin, 2002); the package for calculating nucleosomal potential, which is a measure of DNA affinity for

core histones (Levitsky et al., 2000), http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/; and the package for seeking oligonucleotide motifs differently represented in sequence samples (Vishnevsky, Vityaev, 2001), http://wwwmgs.bionet.nsc.ru/mgs/programs/argo.

Results and Discussion

Chromosome 3

At the first stage of the study, we considered two samples of chromosome 3 sequences differentiated for two properties: PCR^+/PCR^- and HYB^+/HYB^-. The sequences were 2000 bp in length and contained a *Not*I site in the center.

Contrasting oligonucleotide motifs reliably represented in the sample of positive sequences and underrepresented in the negative sample were sought for with the ARGO program (Vishnevsky, Vityaev, 2001). The motifs were sought for in a 50 bp window sliding over the sample with a step of 25 bp. Antagonistic samples were used as negative ones. For example, the PCR^- sample was negative for PCR^+ and vice versa. Thus, four comparison variants were obtained. The proportion of the motifs in the negative sample was no more than 5%. The significance of their presence in the positive sample (probability of random occurrence) was less than 10^{-8}. A 15 single letter–based nucleotide code was used. The length of the motifs was 8 bp. Distributions of the number of significant motifs for each window of each comparison variant were calculated. The well-hybridizing sequences (HYB^+) differed from poorly hybridized ones (HYB^-) in having a peak near nucleotide 570: there were 50 nonrandom octanucleotide motifs virtually absent from HYB^-. One more peak of 40 motifs occurred near nucleotide 1600. Moreover, PCR^- sequences had 41 significant motifs not found in PCR^+ near nucleotide 550. Both pairs showed notable differences at the starts of the sequences.

The nucleosomal potential of all sequences (Levitsky et al., 2000), which characterizes the ability of the DNA stretch under study to form nucleosomes, had a global minimum near nucleotide 1000 (*Not*I site).

Unfortunately, there were too few sequences (21 in the PCR^+ sample and 22 in the HYB^- sample) to construct a significant function discriminating the (+) and (–) classes.

Chromosome 21

Totally 116 sequences of chromosome 21 with a length of 20,000 bp each were subdivided into two samples according to "clonability" (42 $Clon^+$ and 74 $Clon^-$ sequences. Each sequence contained a *Not*I site in the center.

We found that both samples had a gradient of nucleosomal potential along chromosome 21 (Fig. 1).

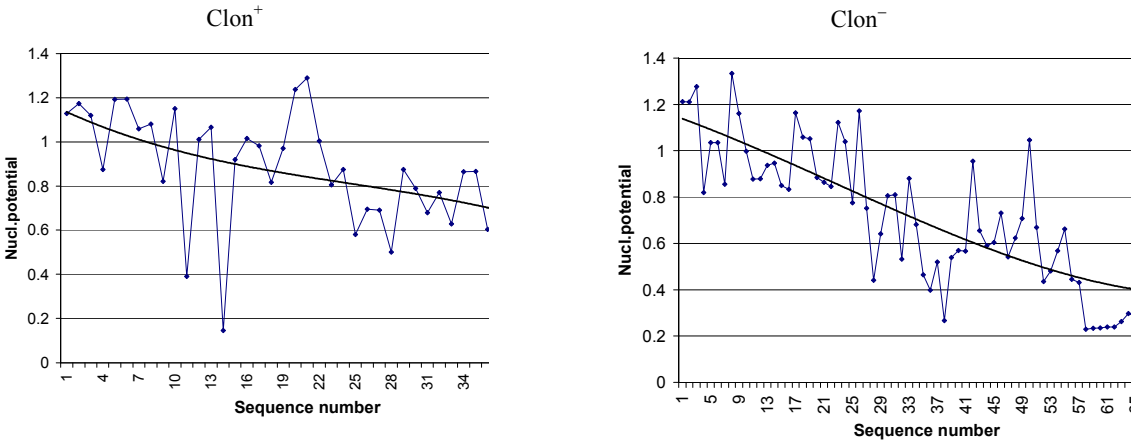


Fig. 1.

All sequences of chromosome 21 have a global minimum of nucleosomal potential in the center, near the *Not*I site (Fig. 2). The samples $Clon^+$ and $Clon^-$ differ dramatically (significance >99.99%) in the local complementation index (LCI) in the range of 4000–4300 bp from the 5'-end or 5700–6000 downstream of the *Not*I site (Likhoshvai, Matushkin, 2002). The sample $Clon^+$ has there much less potential secondary structures than $Clon^-$ (Fig. 3).

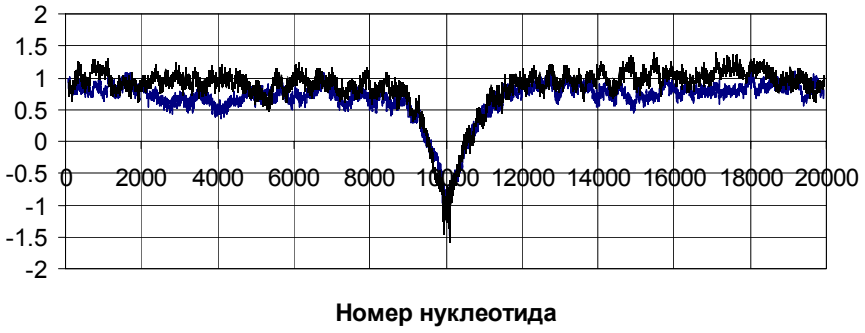


Fig. 2. Chromosomal potential in the $Clon^+$ and $Clon^-$ sequences of chromosome 21. The *Not*I site is located near nucleotide 10,000.

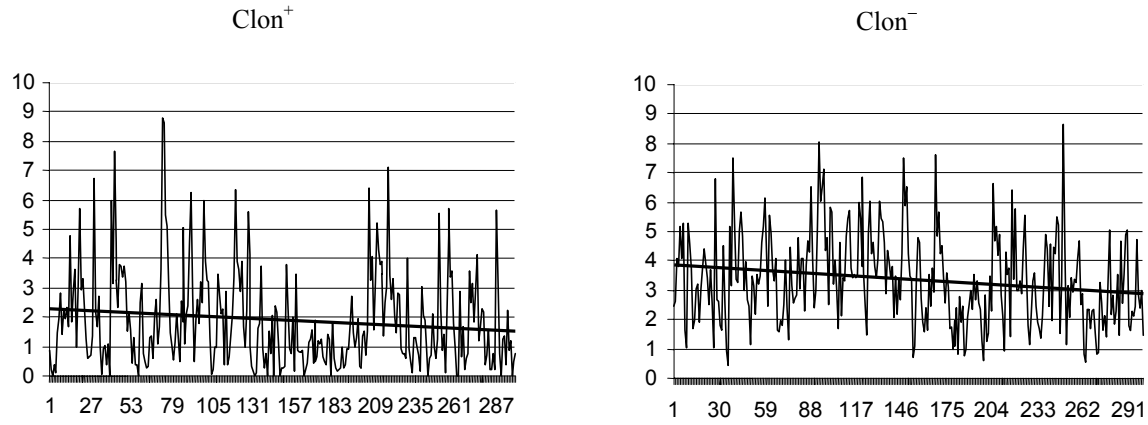


Fig. 3. Local complementation index (LCI). X axis: number of the leftmost nucleotide of the window for which LCI is calculated.

A discriminating function for assignment a sequence to the $Clon^+$ or $Clon^-$ sample on the basis of its specific oligonucleotide pattern was constructed by using the ARGO program (Fig. 4).

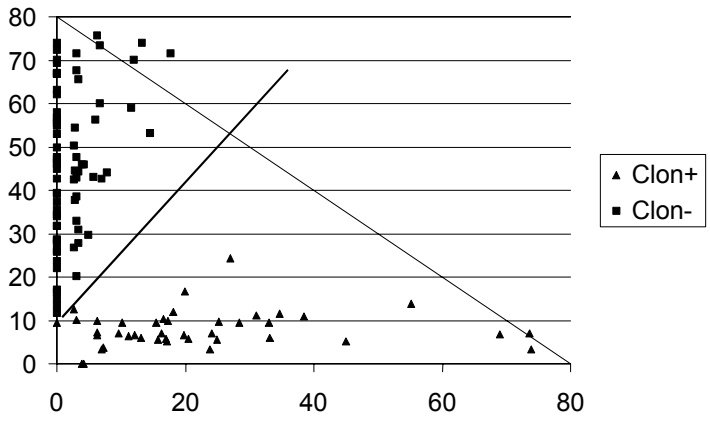


Fig. 4. Values of the discriminating function for all sequences. The diagonal drawn at an angle of 45° strictly separates the two sequence sets.

Chromosome 22

All the methods and algorithms described above were used for investigating 268 sequences of human chromosome 22. One of the tasks was to differentiate them according to "clonability". The sequences were divided into three groups according to the discriminating function constructed using the sequences of chromosome 21. Of these sequences, 38 fell to the group $Clon^+$, 193 sequences, to $Clon^-$, and 37 sequences fell to the indefinite group. The $Clon^+$ and $Clon^-$ groups predicted by ARGO differed significantly in LCI in the range from –100 to +400 with respect to the *Not*I site, which is located in the centers of the sequences.

However, this subdivision does not agree with experimental data, according to which the group $Clon^+$ (Alive) contains 108 sequences; $Clon^-$ (Dead) contains 145; and the indefinite group contains 15 sequences.

There is little agreement between the groups obtained with the use of the discriminating function (calculated from data on chromosome 21) and actual experimental data. Apparently, the sample volumes were insufficient for construction of a universal discriminating function.

The mean nucleosomal potential for each of the 268 sequences in chromosome 22 has no clear trend along the chromosome, contrary to chromosome 21.

Figure 5 shows the distribution of NP values in the region from –100 to +400 bp from the *Not*I site, which occurs in the centers of the sequences, for all 268 sequences of chromosome 22. Nucleosomal potential was calculated for a 160 kb wide window. The proportion of windows with a certain value is shown on Y axis. The general shift to the negative side is indicative of a "pit", which can be seen in Fig. 6.

However, the pit between two distribution humps points to specific selection in this region: the number of structures with NP ≈ –1 is reduced dramatically.

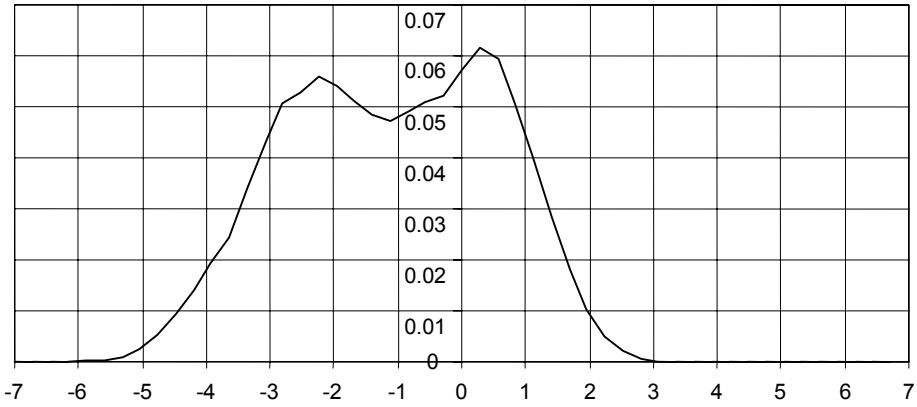


Fig. 5. Distribution of nucleosomal potential in the region from –100 to +400 bp from the *Not*I site for 268 sequences of chromosome 22. X axis: NP; Y axis: the proportion of 160 bp long sequences with this NP value.

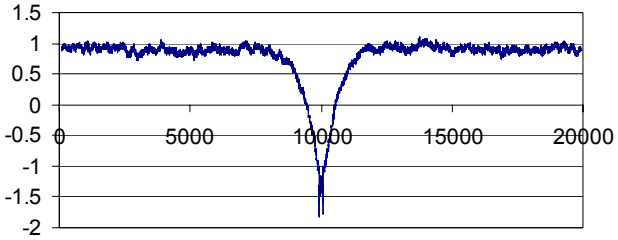


Fig. 6. The profile of NP values (X axis: nucleotide number) averaged over 268 sequences of chromosome 22.

The profile of NP values has a minimum near the *Not*I site for all the studied sequences. Figure 6 illustrates it for chromosome 22. Generally, an NP minimum is characteristic of promoters and like regulatory regions of the genome. Unlike NP, the LCI profile has a maximum at the *Not*I site for all the studied sequences of the human genome. Figure 7 illustrates it for chromosome 22. Hence, secondary structures of high energy are most abundant in this region, but only those that do not favor nucleosome formation.

Thus, *Not*I sites of the human genome are associated with secondary structures minimizing the nucleosomal potential. This is not related to the GC-enrichment of this region, as other GC-rich regions in these sequences do not show such a decrease in NP. The trait of "clonability" for sequences of chromosome 21 is clearly associated with the 300 bp region occurring at a distance 5700–6000 downstream of the *Not*I site.

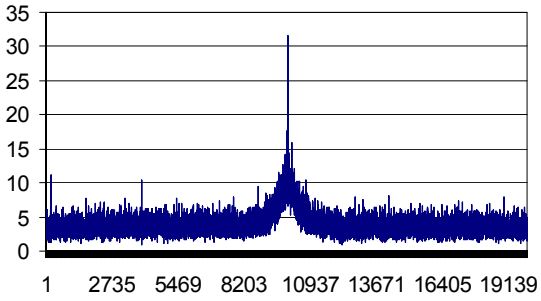


Fig. 7. Average local complementation index along all 268 sequences of chromosome 22. The peak in the neighborhood of the *Not*I site is present in all the studied sequences of the human genome. X axis: nucleotide numbers.

It is not known yet how the presence or absence of high-energy secondary structure in this region can affect the "clonability".

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 02-04-488802, 01-07-90376, and 02-07-90359); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Kashuba V.I., Gizatullin R.G., Protopopov A.I., Allikmets R., Korolev S., Li J., Boldog F., Tory K., Zabarovska V.I., Marcsek Z. et al. (1997). FEBS Lett. 419, 181–185.
2. Kutsenko A.S., Gizatullin R.Z., Al-Amin A.N., Wang F., Kvasha S.M., Podowski R.M., Matushkin Yu.G., Gyanchandani A., Muravenko O.V., Levitsky V.G., Kolchanov N.A., Protopopov A.I., Kashuba V.I., Kisselev L.L., Wasserman W., Wahlestedt C., Zabarovsky E.R. (2002). NotI flanking sequences: a tool for gene discovery and verification of the human genome. Nucl. Acids Res. (In press).
3. Levitsky V.G., Podkolodnaya O.A., Kolchanov N.A., Podkolodny N.L. (2001). Nucleosome formation potential of eukaryotic DNA: tools for calculation and promoters analysis. Bioinformatics. 17, 998–1010.
4. Li J., Wang F., Kashuba V., Wahlestedt C., Zabarovsky E.R. (2001). Biotechniques. 31, 788–793.
5. Likhoshvai V.A., Matushkin Yu.G. (2002). Differentiation of single-cell organisms according to elongation stages crucial for gene expression efficacy. FEBS Lett. 516, 87–92
6. Vishnevsky O.V., Vityaev E.E. (2001). Mol. Biol. (Mosk.). 35, 1–9.

FRAGMENTS OF GENE NETWORK OF FLOWER DEVELOPMENT IN *ARABIDOPSIS* UNDER LONG DAY CONDITIONS AND THEIR DESCRIPTION IN THE GENENET SYSTEM

**** Omelyanchuk N.A., Aksenovich A.V.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: nadya@bionet.nsc.ru

Key words: *flower development, gene network, Arabidopsis*

Resume

Motivation: Considerable volume of data has been so far accumulated on expression regulation of the genes involved in the gene network of *Arabidopsis* flower development. The computer system GeneNet provides convenient tools for accumulation, systematization, regular updating, and graphical representation of these data.

Results: The current published data on gene expression regulation in the course of inflorescence and flower development in *Arabidopsis* were input into GeneNet database. Certain fragments of the gene network regulating *Arabidopsis* floral development are described. Analysis of the partial gene networks constructed allowed us to detect there a number of characteristics common for gene networks regulating ontogenesis, such as positive feedback circuits and cassette-type gene activation.

Availability: http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet

Introduction

Arabidopsis long ago became a model genetic object. Sequencing of its genome attracted additional focus of the plant molecular genetic research. In turn, this increased the accumulation rate of the information important for understanding both the corresponding basic aspects and potential genetic engineering applications. Development of the ABC model, implying involvement of three gene classes in determining the development of the four flower organ types (Coen, Meyerowitz, 1991), enhanced considerably the research into floral development of *Arabidopsis.* Further data accumulation brought into being more sophisticated models of floral development (Mendoza, Alvarez-Buylla, 1998; Mendoza et al., 1999; Tchuraev, Galimzyanov, 2001).

The goal of this work was to apply the possibilities provided by GeneNet to accumulation, systematization, regular updating, and graphical representation of the data on regulation of gene expression resulting in *Arabidopsis* flower development.

Methods and Algorithms

GeneNet is a computer system comprising databases on genes, proteins, signals, events, process, and other data resident in the gene network, and a Java program for visualization these data (Ananko et al., 2002). GeneNet is applicable to processes in which regulation of gene expression has been demonstrated by experimental data on transcription, binding of transcription factors is shown in details as well as the resulting changes in gene expression and the following processes. Unfortunately, these levels of regulation of the flower development are not yet clearly understood. Nevertheless, other experimental data giving evidences on regulation of these processes in an indirect manner, such as changes in mRNA expression in mutant and transgenic lines and resulting changes in the development allowed a virtual model of this gene network, although presumably not so evident and comprehensive, but still informative, to be created. All the data involved in constructing this gene network were extracted from publications. We used the entities substance, gene, mRNA, and protein in combination with direct and indirect relations to construct the database in question.

The database contains a statement that "a gene A regulates directly the expression of a gene B" only if it is known that the transcription factor A binds to the regulatory site housed in the promoter or intron of the gene B and this results in a change in the expression of gene B. It is stated that "a gene A regulates indirectly the expression of a gene B" only if the data on at least two types of the experiments listed below are available:

A mutation of the gene A alters the amount of gene B mRNA;

A mutant phenotype of the gene A is corrected by overexpression of gene B;

On the contrary, a mutation of the gene B suppresses (at least, partially) the phenotype arising from the constitutive expression of the gene A;

A constitutive expression of the gene A changes the expression of gene B; or

Use of certain transgenic constructs involving the gene A (GR fusions or fusing gene A to the activation domain from the viral protein VP16) results in the corresponding expression of gene B.

The genes lacking experimental evidences on the regulatory relations listed above were not included into the database. In the case of redundant actions of genes and co-regulators combined with an unknown mechanism of interactions between these genes, the effect of the gene set is depicted in the gene network as the effect of the major gene whose mutation changes expression of the gene regulated. The paper omits references used for gene network construction, as they all are listed with the GeneNet database.

Results and Discussion

When an increase in the daylight induces flowering, the light signal via photoreceptors is transmitted to the system of light signal transduction (Fig.) and to the biological clock system. The last system includes the genes *GIGANTEA* (*GI*) and *LATE ELONGATED HYPOCOTYL* (*LHY*), whose activities are altered under long day conditions, resulting in a change in the activity of gene *CONSTANS* (*CO*). The last gene is a "middleman" between the biological clock system and gene network of floral development. Expression of gene *CO* in the vegetative meristem initiates and maintains the expressions of two genes—*TERMINAL FLOWER1* (*TFL1*) and *LEAFY* (*LFY*). *TFL1* is the major gene detaining the meristem at the vegetative stage through complete inhibition of the gene *AP1* activity and restraining *LFY* to a particular activity level. A long day–stimulated increase in the gene CO activity results in considerable elevation in the activities of two genes—*FT* (*FLOWERING LOCUS T*) and *LFY*—in the cells of peripheral vegetative meristem. *FT* induces the activity of gene *AP1* in an indirect manner. *LFY* encodes a transcription factor and is the major gene in the floral development process. Rapid initial activation of the floral development gene network and transition of flank meristem cells to formation of flower primordia requires considerable amount of the protein *LFY*. This is achieved due to a positive feedback circuit of the genes *LFY* and *APETALA1* (*AP1*). High levels of *LFY* and *APETALA1* proteins also underlie the irreversibility of transition to flowering, as they inhibit the gene *TFL1*, which suppresses the activities of these genes in vegetative meristem. This positive feedback regulatory circuit maintains the gene *LFY*, the major gene of the floral development process, in the state of being "constantly switched on", triggering the cascade of transcription factor genes *AP1, AP3, PI*, and *AG*. These transcription factors provide formation of various flower primodria, namely sepals (*AP1*), petals (*AP1, AP3,* and *PI*), stamens (*AP3, PI*, and *AG*), and carpel (*AG*).

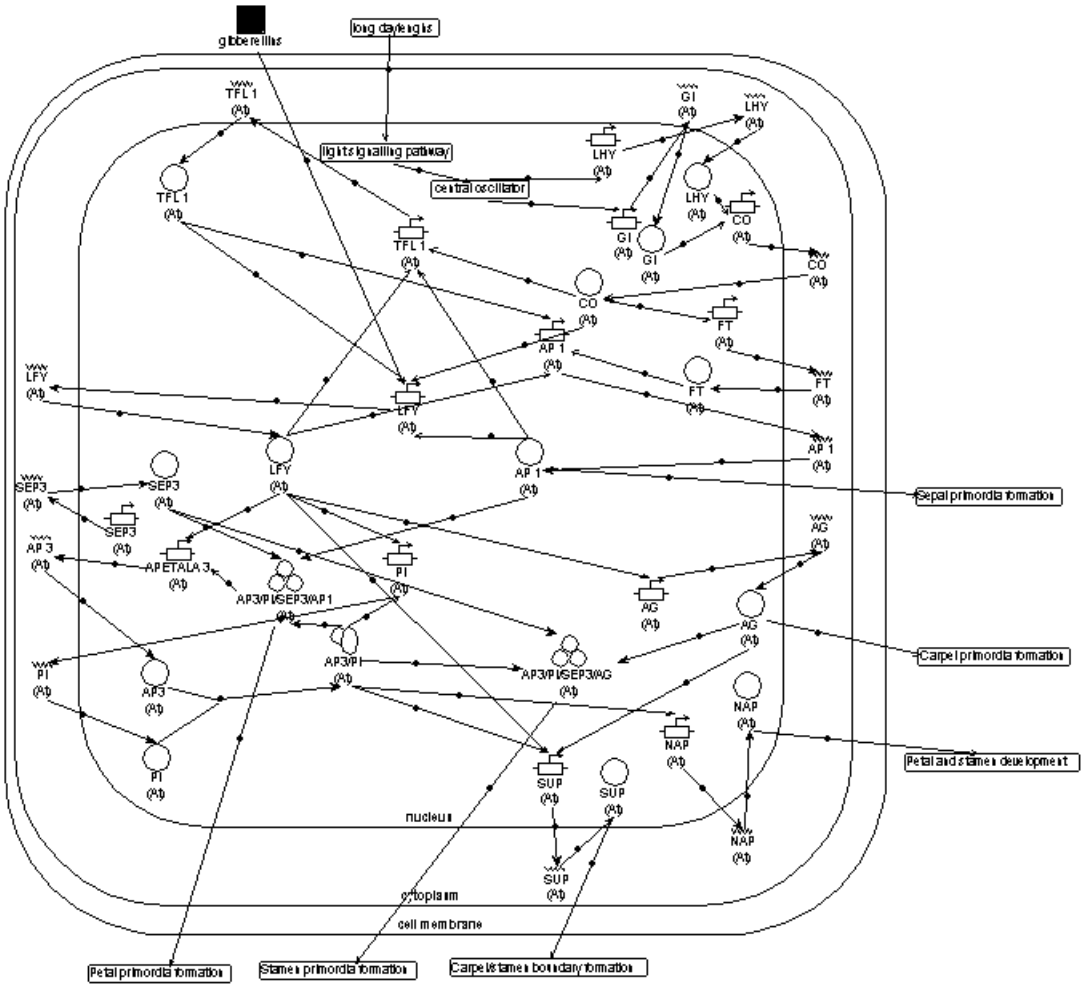


Fig. Fragment of the gene network of flower development in *Arabidopsis*. Filled ovals denote the proteins; filled rectangles, genes; and arrows, regulatory effects.

The major regulator of other significant positive feedback regulatory circuit supplemented with autoregulation is heterodimeric protein AP3/PI, involved in maintaining the transcription of both *AP3* and *PI* genes, coding for its monomers. It has been demonstrated that this heterodimer is a component of the tetramer AP3/PI/SEP3/AP1, which increases the gene *AP3* transcription and is the major factor providing formation of the petal primodria. As for stamen primodria, the key factor here is the tetramer AP3/PI/SEP3/AG. Heterodimer AP3/PI also participate in switching on the genes *SUPERMAN* (*SUP*), responsible for forming the border between stamens and carpels, and *NAP*, providing the further development of stamens and petals.

A graphical layout of the gene network in question makes the relation between the time course of transcription regulation and developmental processes in question more illustrative. The gene network of flower development in *Arabidopsis* has a number of characteristics common for gene networks of developmental processes, such as positive feedback circuits and cassette-type gene activation. Our computer system presents the current experimental data on regulation of gene expression in flower development and provides their updating with new information published on already described and novel genes on a regular basis.

Acknowledgements

This work was supported in part by the Russian Foundation for Basic Research (grant № 00-04-49255). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002). GeneNet: a database on structure and functional organization of gene networks. Nucl. Acids Res. 30, 398-401.

2. Coen E.S., Meyerowitz E.M. (1991). The war of the whorls: genetic interactions controlling flower development. Nature. 353, 31-37.

3. Mendoza L., Alvarez-Buylla E.R. (1998). Dynamics of the genetic regulatory network for Arabidopsis thaliana flower morphogenesis. J. Theor. Biol. 193, 307-319.

4. Mendoza L., Thieffry D., Alvarez-Buylla E.R. (1999). Genetic control of flower morphogenesis in Arabidopsis thaliana: a logical analysis. Bioinformatics. 15, 593-606.

5. Tchuraev R.N., Galimzyanov A.V. (2001). Modeling of actual eukaryotic control gene subnetworks based on the method of generalized threshold models. Mol. Biol. (Mosk.). 35, 1088-1094.

RECOGNIZING FUNCTIONAL DNA SITES AND SEGMENTING GENOMES USING THE PROGRAM "COMPLEXITY"

[1] Orlov Yu.L., [2] Potapov V.N., [1] Filippov V.P.*

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

Key words: *sequence analysis, complexity, data compression, complete genomes, suffix tree visualization, variable memory Markov model*

Resume

Motivation: Large-scale sequencing of genomes opens new opportunities for analysis of long nucleotide sequences. Of special interest is the detection of common contextual properties that are stable to evolutionary changes in terms of the genome. The next problem is segmentation of genome sequences on the basis of these context properties.

Results: A program that constructs Markov models with variable memory for generating genetic texts was developed. Using these variable memory Markov models, a method for recognizing functional DNA sites was developed and tested on promoters containing TATA-box sites. A method for segmentation of genomic sequences was developed using an estimated probability of observing a certain region taking into account its local contexts.

Availability: http://wwwmgs.bionet.nsc.ru/programs/complexity/.

Introduction

Large-scale projects of sequencing of whole microbial and eukaryotic genomes open promising opportunities for comparison of DNA sequences in different organisms. Of special interest is the analysis of most general characteristics of genome sequences. Information measures of symbol sequences exemplify such general characteristics (Haring, Kypr, 1999).

The algorithm proposed, basing on methods of the data compression theory, allows construction of the variable memory Markov model to generate sequences (Orlov, Potapov, 2000; Orlov et al., 2002). The model of text generation is represented unambiguously by suffix trees (Ron et al., 1996).

"Complexity", an Internet-accessible software tool (http://wwwmgs.bionet.nsc.ru/programs/complexity/), generates probabilistic suffix trees (PST) for a specified nucleotide sequence. The representation of the model as a generation tree source in a GIF format facilitates visual comparison of the inner structure of the texts.

Methods and Algorithms

Let us consider a stationary stochastic grammar model of text generation. The model T generates the sequence $X^n = X_1X_2...X_n$ with a probability: $P(X^n) = P(X_1|S_1)P(X_2|S_2)...P(X_n|S_n)$. The probability $P(X_n|S_n)$, which is independent of the position of n in the sequence, is determined only by the preceding context S_n. The probability $P(D_i|S_j)$ of occurrence of a certain letter $D_i \in \{A,T,G,C\}$ $(i = 1,2,3,4)$ in each of the contexts S_j is determined as follows by the corresponding parameters of the distribution θ:

$P(D_i|S_j)=\theta^i_j$,

where $\sum_{i=1}^{4} \theta^i_j = 1$, $j = 1,2,...|T|$, $|T|$ is the total number of contexts in the model T (number of leaves in the tree), and $4 \le |T| \le 4^k$ (k is the maximum context length).

Equivalent states of the Markov model corresponding to contexts of various lengths can be integrated using the algorithm developed for constructing contextual probabilistic tree sources (Orlov, Potapov, 2002; Orlov et al., 2002), which employs algorithms previously developed within the source coding and data compression theories (Barron et al., 1998).

A context tree can be represented graphically; the contexts might vary in length, and none of them is the end of another (Fig. 1).

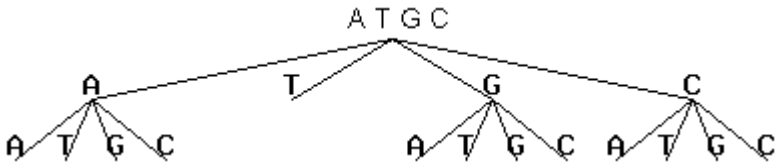


Fig. 1. Example of a generating tree source. The tree is constructed using the program Complexity for nucleotide sequences of the AP-1 transcription factor binding site (http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/). Each link from the leaves to the root corresponds to a context in a DNA sequence and has its own set of probabilities of generating the next symbol.

For example, in the tree T shown in Fig. 1, there are twelve contexts with a length of two nucleotides (AA, TA, GA, AG, TG, GG, CG, AC, TC, GC, and CC) and one context with a length of one nucleotide (T). Totally, there are 13 preceding contexts, i.e., $|T| = 13$. Each context specifies four numbers: the probabilities of generating symbols located to the right of this context. To determine these numbers, we use frequencies of the corresponding oligonucleotides that are by one nucleotide longer: totally, we have $4 \times 13 = 52$ numbers.

Implementation and Results

Construction of generating models for extended DNA sequences

Earlier, samples of nucleotide sequences of several functional classes were analyzed using the method proposed (Orlov, Potapov, 2000). Analysis of the nucleotide sequences of various functional classes (coding, noncoding, and regulatory regions) shows that DNA sequences have tree sources of various structures. The models differ in the order of the Markov chain, the tree structure, and number of branches (Orlov et al., 2002) (see Figs. 1 and 2). Figure 2 shows that models can be more complex than a model shown in Fig. 1.

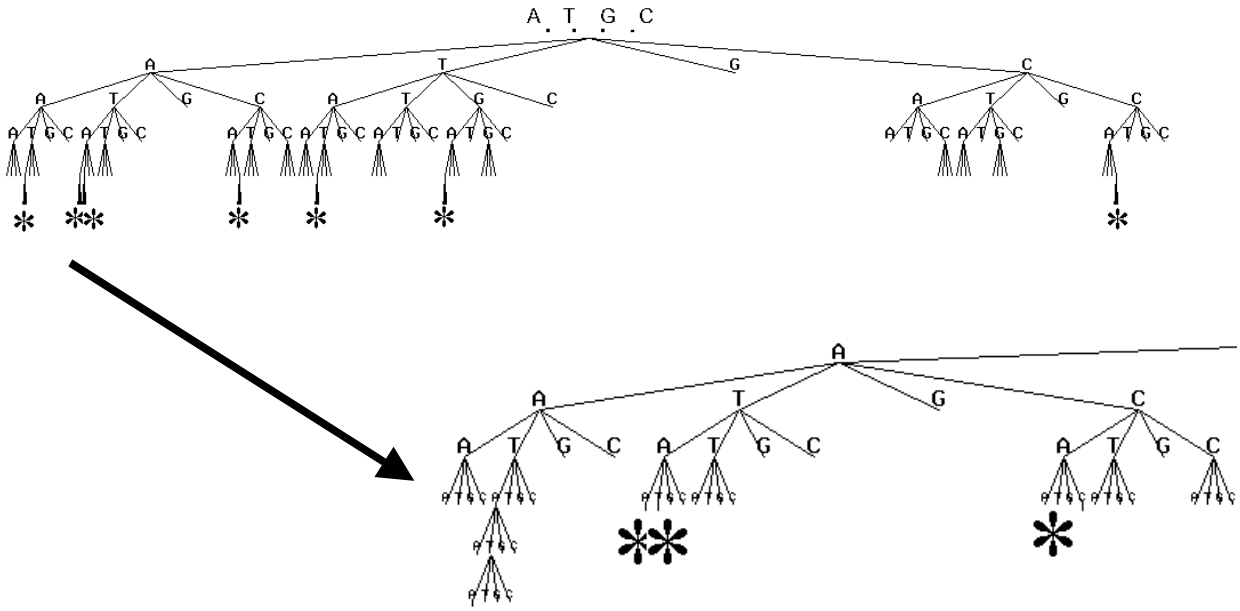


Fig. 2. Context tree source for nucleotide sequences containing TATA box. (534 sequences were extracted from the TRRD database, the sequence length was 20 bp). The letters for preceding contexts with a length of more than 3 bp are not shown because of a limited space in the scheme; an asterisk (*) denotes contexts more than 5 bp in length. Below, we give an enlarged fragment of the same tree, which is bordered with a gray line. Long contexts marked with the first asterisk from the left are shown in full: ATATAA, TTATAA, GTATAA, and CTATAA.

There is a technical problem of a simultaneous graphic representation of all contexts. It is impossible to arrange contexts of a length of more than 3–4 nucleotides (totally, 64–256 possible contexts) on a standard page or on a computer screen; therefore, an image is given iteratively.

The model developed allows the assessment of the degree to which certain sequences are close to a specified sequence in terms of their local contexts. These assessments can be used to (1) recognize relatively short regions in long sequences and (2) segment long DNA sequences into regions that differ in their contextual compositions.

Recognition of functional sites

Let us consider the problem of recognition of functional regions in extended sequences. In this case, the model is constructed not for one sequence but for a sample of functional regions. For a sample of sequences, a probabilistic tree source with a set of probabilities of generating symbols from the DNA alphabet is calculated taking into account that several first leftward symbols forming the first context are missing. A tree source was constructed for the sample of TATA-box sequences from the TRRD database (see Fig. 2). Then, promoter sequences from the TRRD database were analyzed. The probability of obtaining this region was calculated by a sliding window of 20 bp and the logarithmic profile for such a probability with the minus sign was constructed (Fig. 3). The minimal value of the profile is in the region with TATA boxes that are most typical of this model. The minimal profile value for the metallothionein-I gene promoter (AC EMBL: J00605) corresponds to an actual TATA box indexed in the TRRD database at a distance of –28 to –23 bp before the transcription start (denoted by the arrow in Fig. 3).

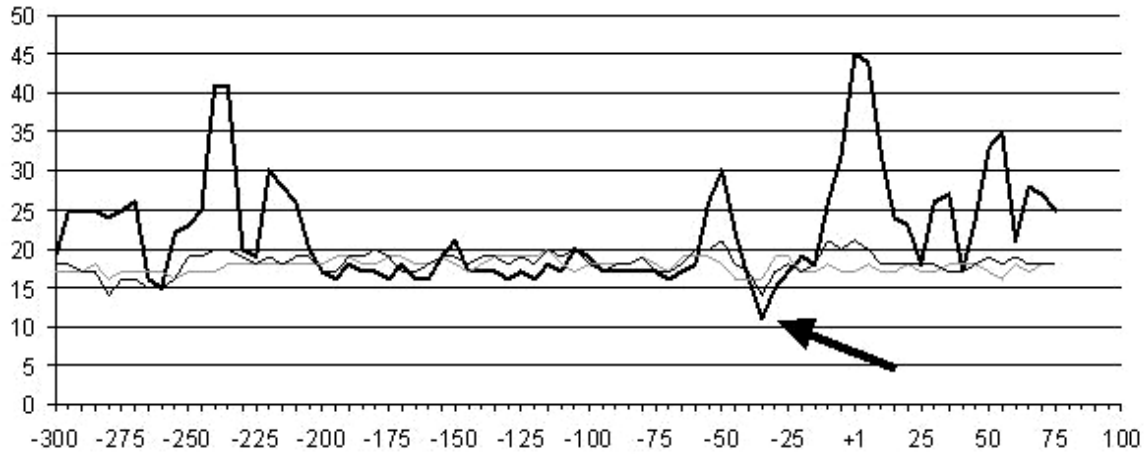


Fig. 3. Profile of the TATA box recognizing function in the promoter region [-300;+100] of the metallothionein-I gene. The solid curve refers to recognition in the tree source model. The gray and thin black curves refer to recognition by the nucleotide and dinucleotide frequencies in the same sliding window.

As is evident, the profile of the recognition function in the variable memory Markov model shows a more prominent TATA-box region in comparison with that recognized by the standard Makov models using only nucleotide and dinucleotide frequencies.

In constructing a model of a generating tree source for a corresponding sample of functional sites, it is unnecessary to align preliminarily the sites or even determine their boundaries. Positioning of symbols is not necessary in contrast to the weight matrix. Local contexts substitute positioning. Thus, this method of simulation is an appropriate alternative to the weight matrix method.

Genome segmentation

Let us consider the problem of segmentation of long sequences. The variable memory model allows recognition of typical and atypical regions of the entire sequence, i.e., the latter are the regions with a low probability of being accidentally observed. The probability for a region is calculated as a probability of observation of all the symbols taking into account the previous local context. In this case, the training involves the sequences of the entire genome.

The probability profile constructed by a sliding window of a length of 1 kbp shows notable regions that are statistically atypical of the entire genome. Figure 4 gives an example of a profile in a sliding window for the *Bacillus subtilis* genomic sequence. The arrows show the regions that are most atypical of the genome. These are regions of ribosomal protein genes (14 kbp), ribosomal and transport RNAs (98–100 kbp, 165–170 kbp, 635 kbp, 946–950 kbp, and 3,172 kbp, respectively). Thus, the regions that belong to the translation system—the most ancient and conservative system in living organisms—differ mostly in the local context composition.

The position (bp) is plotted on the X axis, and a sum of logarithms of the observation probability for nucleotides (with the minus sign) in a 1 kbp window of the tree source model is plotted on the Y axis.

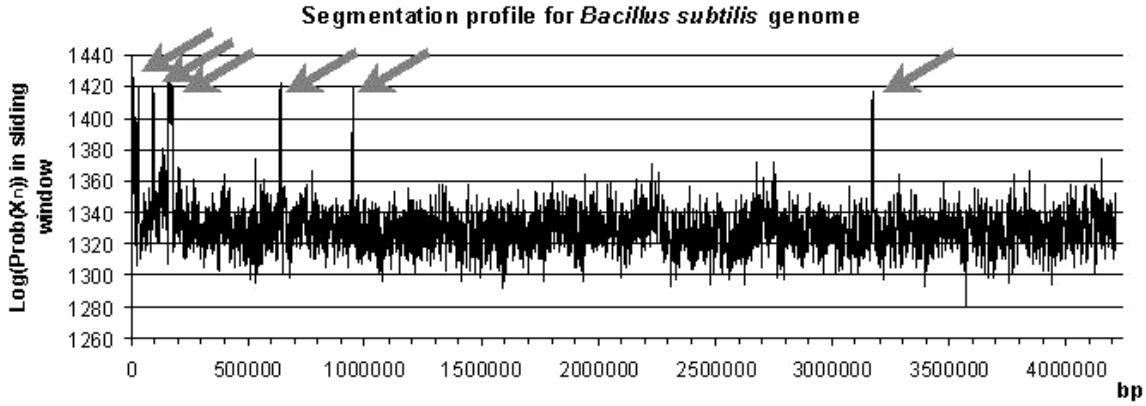


Fig. 4. Profile of compliance with the general model (statistical proximity) of local regions of *Bacillus subtilis* genome in a sliding 1 kbp window.

Discussion

The program Complexity allows us to construct a model of generation of a genetic text and determine the complexity of this text. The presented version of the program is designed for analysis of long DNA sequences of any size.

An important characteristic of genomic sequences is their oligonucleotide composition, which reflects evolutionary interactions between organisms (Karlin, Ladunga, 1994; Scherer et al., 1994). Examples of distribution of short oligonucleotides and clustering of significant oligonucleotides were studied by Haring and Kypr (1999). However, a unified method for the representation of oligonucleotides specific of a genomic DNA has not been developed. A graphic representation in the form of a tree structure is rather illustrative. The program developed presents a tree structure of oligonucleotides based on the selection by the minimum description length principle (Barron et al., 1998). This distinguishes our model from a graphic representation of statistically under-represented and over-represented oligonucleotides constructed by the program Verbumculus (http://wwwdbl.dei.unipd.it/Verbumculus/) (Apostolico et al., 2000).

The recognition of functional regions using variable memory Markov models is a promising method. Currently, we work on segmentation of all sequenced genomic sequences.

Acknowledgements

The authors thank V.A.Likhoshvai, M.A.Pozdnyakov, and N.A.Kolchanov for their helpful discussions. The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 00-07-90337, 02-07-90355, 00-04-49229, and 02-01-00939); Ministry of Industry, Science, and Technologies of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); and the INTAS foundation (grant № YSF 00-178).

References

1. Apostolico A., Bock M.E., Lonardi S., Xu X. (2000). Efficient detection of unusual words. J. Comput. Biol. 7(1/2):71–94.
2. Barron A., Rissanen J., Yu B. (1998). The minimum description length principle in coding and modeling. IEEE Trans. Inform. Theory. 44:2743–2760.
3. Haring D., Kypr J. (1999). Variations of the mononucleotide and short oligonucleotide distributions in the genomes of various organisms. J. Theor. Biol. 201:141–156.
4. Karlin S., Ladunga I. (1994). Comparisons of eukaryotic genomic sequences. Proc. Natl Acad. Sci. USA. 91:12832–12836.
5. Orlov Yu.L., Filippov V.P., Potapov V.N., Kolchanov N.A. (2002). Construction of stochastic context trees for genetic texts. (Bioinformation Systems e.V.) In Silico Biology, 2(0022) <http://www.bioinfo.de/isb/2002/02/0022/>
6. Orlov Yu.L., Potapov V.N. (2000). Estimation of stochastic complexity of genetic texts. Computational Technologies (Novosibirsk), 5:5–15.
7. Ron D., Singer Y., Tishby N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. Machine Learning. 25:117–149.
8. Scherer S., McPeek M.S., Speed T.P. (1994). Atypical regions in large genomic DNA sequences. Proc. Natl Acad. Sci. USA. 91:7134–7138.

SOFTWARE PACKAGE LZCOMPOSER: ANALYSIS OF OCCURRENCE OF REPEATS IN COMPLETE GENOMES

[1]* Orlov Yu.L., [2] Gusev V.D., [2] Nemytikova L.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Mathematics, SB RAS, Novosibirsk, Russia
*Corresponding author. e-mail: orlov@bionet.nsc.ru

Key words: *Lempel-Ziv complexity, direct and inverted repeats, complete genomes, genome segmentation*

Resume

Motivation: Complexity of a genome may be estimated as the number of duplications, inversions, and DNA rearrangements, which are related to evolutionary history. Detection of direct and complementary repeats in the sequences of complete genomes gives a basis for analysis of genome structure.

Results: By using the software package LZcomposer, we have analyzed the structure of repeats at the genome scale and demonstrated the statistical significance of direct and inverted repeats.

Availability: The software package LZcomposer for complexity decomposition is available via the Internet by the address http://wwwmgs.bionet.nsc.ru/programs/lzcomposer/.

Introduction

For studying inherited information, it is of interest to evaluate entropy parameters, as well as to search for perfect (completely coinciding) and imperfect repeats. Decomposition of nucleotide sequence into non-overlapping fragments, each with its own "prototype", that is, direct or complementary repeat of maximal length in the other textual parts, enables to represent clearly the structure of the sequence analyzed, to detect highly repeated sequences (Gusev et al., 1991; see also Gusev et al., 2002, this issue). However, despite there exist some programs aimed at searching for repeats and their visualization (Kurtz et al., 2001), the general estimation of genome complexity as a measure of saturation by repeats was not done. Probably, this fact could be explained by computational difficulties and heterogeneity of the tasks that need analysis of repeats, in particular, for segmentation of genomes, estimation of phylogenetic relationships, comparison with EST-sequences. In particular, repeats may either intersect, or more extended but imperfect repeats may be more significant than short perfect ones. We apply suggested by V.D.Gusev and co-authors estimation of complexity for generation of the text by Lempel and Ziv, as the minimal number of copy operations necessary for reproducing a sequence on the basis of this sequence itself (Gusev et al., 1991). This methodology of complexity estimation was applied for analyzing relatively short DNA sequences, up to 100 000 bp in length (Gusev et al., 1999; Gusev et al., 2001), and it is well suitable for searching for local repeated regions. The program software LZcomposer (http://wwwmgs.bionet.nsc.ru/programs/lzcomposer/) calculates complexity of extended sequences as a whole, down to the size of a complete bacterial genome or eukaryotic chromosomes. In the present work, we study distributions of lengths of various types of repeats.

Methods and Algorithms

Algorithm of calculating DNA sequence complexity by using one of preordered operations, namely, direct, symmetric, direct complementary, and inverted (symmetric complementary) copying is described in details in (Gusev et al., 1999; Gusev et al., 2001). Thus, the algorithm presents a generalization of Lempel-Ziv scheme. The program fulfils consequent (from left to the right) decomposition of the nucleotide sequence into non-overlapping fragments, each possessing by its own "prototype", which is a direct or complementary repeat of maximal possible length in preceding part of the text. Classification of types of repeats is given in Figure 1 in (Gusev et al., 2002, this issue). Additionally, the data are displayed about the average, minimal, and maximal lengths of components of decomposition, as well as statistics estimating occurrence of repeats of different types. The histograms of complexity values shown in the windows with the fixed length are also of specific interest, because they enable to compare both the real genome with its "random" copies and genomes of different organisms if their genome sizes are normalized by length.

Implementation and Results

Comparing complexity decompositions of nucleotide sequences is of interest for different organisms, by using at most complete data. To this aim, we have analyzed complete genome sequences of several bacterial genomes extracted from GenBank.

Distribution of number of repeats in *Archaeoglobus fulgidus, Aquifex aeolicus,* and *Borrelia burgdorferi* genomes in relation to repeat length is given in Figure 1.

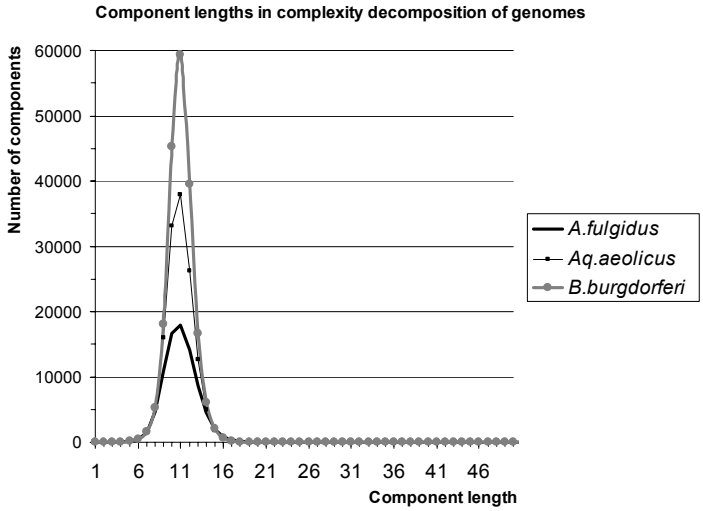


Fig. 1. The number of components of complete complexity decomposition of genome sequences in accordance with the length of components (allowance for all 4 types of repeats was included). By abscissa, the length of a component is given, whereas by ordinate, the total number of such components in complete decomposition. Genome sizes in *Archaeoglobus fulgidus, Aquifex aeolicus,* and *Borrelia burgdorferi* equal to 2178470, 1551400, and 910748 bp, respectively. Peak height depends upon genome size, distribution mode is constant.

An average repeat length in complexity decomposition varies from 10 to 13 nucleotides not only for those repeats illustrated in Figure 1, but also for all the rest genomes analyzed (see Table). Notably, the most frequent length value is rather stable and equals to 12 nucleotides in all bacterial genomes studied. The distribution tail area shown in Figure 1 is markedly "overextended": in some genomes, there are repeats of several thousands bp in length. It should be noted that the length of the maximal repeat is rather unstable parameter, which is determined not only by the length, but also by specificity of repeats, i.e., extended end repeats, tandem repeats with high multiplicity, inserted mobile elements, etc.

Table. Complexity decompositions parameters of bacterial genomes and eukaryotic nucleotide sequences.

Organism	Genome size (bp)	Average length of decomposition components*	Maximal component length and type **	Average component length, calculated only for:			
				Standard repeat		Complementary repeat	
				Direct	Symmetry	Inverted	Complementary
Mycoplasma pneumoniae	816394	10.86	459 D	12.01	10.23	10.56	10.23
Borrelia burgdorferi	910724	11.02	1430 D	11.13	10.90	11.05	10.95
Plasmodium falciparum chr.2	947089	12.99	475 I	13.83	12.23	13.33	12.22
Chlamydia trachomatis	1042519	10.45	4904 D	10.64	10.33	10.44	10.37
Rickettsia prowazekii	1111523	10.99	487 D	11.04	10.93	11.01	10.95
Treponema pallidum	1137944	10.50	3278 D	10.84	10.28	10.47	10.31
Chlamydia pneumoniae	1230230	10.59	1371 D	10.77	10.47	10.59	10.52
Aquifex aeolicus	1551335	11.00	5270 I	11.10	10.80	11.16	10.84
Methanococcus jannaschii	1664957	11.56	1018 I	11.66	11.33	11.73	11.34
Helicobacter pylori 26695	1667825	11.44	4847 I	11.67	10.93	11.70	10.93
M. thermoautotrophicum	1751377	10.96	1856 D	11.28	10.61	10.99	10.64
Haemophilus influenzae Rd	1830023	11.28	5791 I	11.50	10.89	11.61	10.91
Thermotoga maritima	1860725	11.01	917 D	11.17	10.80	11.12	10.79
Archaeoglobus fulgidus	2178400	11.11	1209 D	11.37	10.83	11.21	10.85
Synechocystis PCC6803	3573470	11.57	5353 I	11.82	11.18	11.82	11.20
Mycobacterium tuberculosis	4411529	12.08	1697 D	12.41	11.72	12.20	11.72
Average	-	11.21	-	11,51	10,90	11,31	10,92
Eukaryotes:							
Mouse chromosome 1	4278458	13.42	10187 I	14.32	11.51	14.63	11.48
Mouse hromosome 2	3438198	12.15	43688 I	13.11	11.40	12.04	11.36
Mouse hromosome Y	369611	10.74	10095 I	11.02	9.85	11.70	9.71
S. cerevisiae Chromosome I	230205	9.71	810 I	9.75	9.15	10.62	9.19
S. cerevisiae Chrom. VIII	562641	10.10	1988 D	10.29	9.90	10.26	9.92
Computer simulation (equal frequencies of nucleotides):							
Random sequence	1000000	10.10	21	9.09	9.09	9.10	9.08

Note: *In the third column, we accounted for repeats referring to all 4 types (bold-typed). The length of the maximal component, in general case, does not coincide with the length of the maximal repeat, but it is a good lower estimate. ** Type of a repeat: D - Direct, I -Inverted.

The main parameters of complexity decompositions of DNA sequences for genomes of various microorganisms are shown in Table. Sequence complexity that equals to the quotient of genome size dividing by the average length of decomposition components is not shown in the Table, but it could be easily calculated. The primary interest is caused not by complexity as it is, but by the most informative decomposition components (extra-long repeats, their types, mutual disposition, expansibility by taking into account admissible distortions, etc.).

Table 1 demonstrates that the average component length in complexity decomposition with accounting for all types of repeats equals approximately to 11 nucleotides, whereas for decomposition only for direct repeats, to 10 nucleotides. In other words, the choice of optimal (i.e., maximal by length) type of repeat at every step of decomposition increases average component length for one symbol. As known, the average repeat length is higher in eukaryotes than in prokaryotes, this fact being supported by comparison of average values for mouse chromosomes and bacterial genomes (Table). Also, the qualitative estimation of beneficial usage of direct and inverted repeats stays the same.

To compare the differences in frequencies of occurrence of repeats of different types, we have made a computer simulation experiment with random sequences. As shown for random sequences with equaling nucleotide frequencies, with the length of 910724 nucleotides, corresponding to genome size of *Borrelia burgdorferi*, the average lengths of repeats of all types are equal and correspond to 10 nucleotides (see Table, the raw "Computer simulation"). Generation of several random sequences of the same length and the same nucleotide content as those in *Borrelia burgdorferi* genome gives slightly higher average value of direct and symmetric repeat length, 10.5 nucleotides. However, the difference between direct and symmetric repeats was not found. Notably, repeats with the length exceeding 21 bp were not found in random sequences. Thus, by computer simulation experiment, we have demonstrated significance and non-randomness of direct and inverted repeats.

By using all types of repeats, we arrive at the most compact (with lesser number of components) decomposition. For each particular type of decomposition, it is interesting to know the balance between occurrences of repeats of different types. Direct repeats are always more frequent, next follow inverted (symmetric complementary) repeats (hairpin structures, complementary palindromes), and then symmetric and direct complementary repeats (see classification of repeats in Gusev et al., 2002, this issue).

Such ordering reveals origin of repeats in evolution, which is related to duplications and inversions. This fact is supported by analysis of complexity decompositions based on usage of repeats referring only to a single type (see the last four table columns). The similar pattern in occurrence of repeats of different types in promoter regions was previously demonstrated in (Babenko et al., 1999). Also, it is of interest to study occurrences of repeats with the length exceeding 10 nucleotides, as well as distribution of maximal values in this case. As was shown, more extended symmetric and direct complementary repeats (with the length at least 30 bp) are absent in complexity decompositions within genome scale. All long repeats were either direct, or inverted (data not shown in the Table). Direct repeats with the length exceeding 21 bp are a little bit more frequent than inverted repeats of the same length (see Fig. 2).

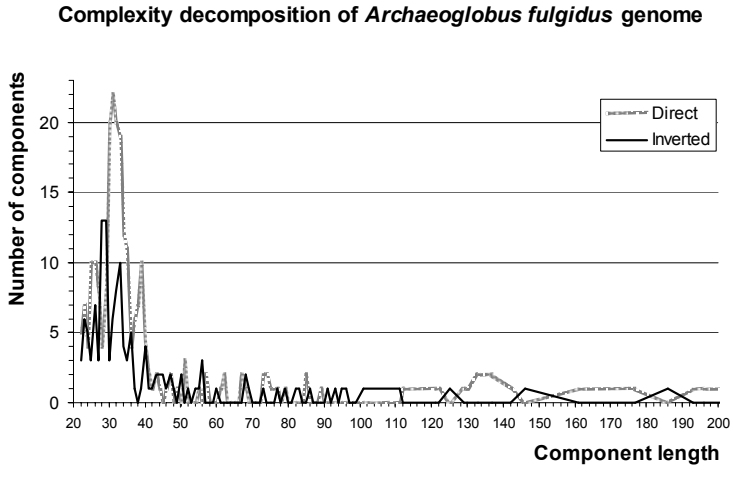


Fig. 2. Number of components of complete complexity decomposition of *Archaeoglobus fulgidus* genome in accordance with component length. The data are shown for repeats with the length at least 21 bp. By gray line, direct repeats are shown; by black line, inverted repeats (for details of scaling, see Figure 1). In decomposition, we have considered the repeats of all four types, however, symmetric and direct complementary repeats were never met. Repeats with maximal lengths are not shown on the plot. The maximal repeat, of 1209 bp in length, was direct.

Discussion

Analysis of extended genome sequences (of the order of 10^6 bp and higher) have revealed their saturation by repeats, primarily by direct and direct complementary (inverted) ones. Direct and inverted repeats originate in evolution as a result of duplications and duplications with subsequent inversions. As for symmetric and direct complementary repeats, the analogous genetic mechanisms for their generation are unknown. Such repeats may appear as a result of convergent evolution, so their length is determined only by combinatorial factors, as we observe in reality (Table). However, for analysis of local DNA regions, for example, regulatory regions, symmetric and complementary repeats may be significant

due to DNA-protein interactions. Such analysis of complexity in promoter regions was made by the software program "Complexity Profile Builder" (http://wwwmgs.bionet.nsc.ru/mgs/programs/gc_net/), which realizes an algorithm for the sequences with the lengths up to 32 Kb (Babenko et al., 1999).

We have demonstrated the pattern of usage of repeats within the genome scale. The same methodology of complexity decompositions provides rather convenient standard for evaluation of complexity of sequences. This approach is applicable both within the scale of 100-1000 bp for analysis of regulatory regions (Babenko et al., 1999) and within the scale 10-100 bp, for analysis of mutations (Krawczak et al., 2000).

As shown, the repeats with the length of 10-11 nucleotides are predominant in sequence decompositions (Fig. 1). This fact is in a good agreement with estimations of statistical significance of oligonucleotides with corresponding lengths of 10-11 nucleotides in the course of constructing context tree source models for the same genomes (Orlov et al., 2002).

Acknowledgements

The authors are grateful to V.A.Likhoshvai for assistance in preparing the complete genome sequences for computations, to V.P.Filippov for programming, to N.A.Kolchanov for fruitful discussions of results. This work was supported in part by the RFBR (grants № 01-07-90376, 00-07-90337, 02-07-90355, 00-04-49229, 02-01-00939, 00-06-80420), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Project № 65). Y.O. was supported by INTAS (YSF 00-178).

References

1. Babenko V.N., Kosarev P.S., Vishnevsky O.V., Levitsky V.G., Basin V.V., Frolov A.S. (1999) Investigating extended regulatory regions of genomic DNA sequences. Bioinformatics. 15(7-8), 644-53.
2. Gusev V.D., Kulichkov V.A., Chupakhina O.M. (1991) Complexity analysis of genomes. Measures of complexity and classification of the structural regulations revealed. Mol. Biol. (Mosk.). 25, 825-834. (In Russ.).
3. Gusev V.D., Nemytikova L.A. (2001) Calculation of repeatedness, symmetry and isomorphism for symbolic sequences. Methods for searching for empirical regularities. (Informational systems, Novosibirsk, 167), 11-33. (In Russ.).
4. Gusev V.D., Nemytikova L.A., Chuzhanova N.A. (1999) On the complexity measures of genetic sequences. Bioinformatics. 15(12), 994-999.
5. Gusev V.D., Nemytikova L.A., Chuzhanova N.A. (2001) A rapid method for detecting interconnections between functionally and/or evolutionary close biological sequences. Mol. Biol. (Mosk.). 35(6), 1015-1022. (In Russ.).
6. Gusev V.D., Nemytikova L.A., Orlov Yu.L., Filippov V.P. (2002) Internet-available software package LZcomposer: analysis of structure of genome sequences by complexity decomposition (this issue).
7. Krawczak M., Chuzhanova N.A., Stenson P., Johansen B., Ball E., Cooper D.N. (2000) Changes in primary DNA sequence complexity influence the phenotypic consequences of mutations in human gene regulatory regions. Hum. Genet. 107, 362-365.
8. Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucl. Acids Res. 29(22), 4633-42.
9. Orlov Yu.L., Filippov V.P., Potapov V.N., Kolchanov N.A. (2002) Construction of stochastic context trees for genetic texts. (Bioinformation Systems e.V.) In Silico Biology 2, 0022 (http://www.bioinfo.de/isb/2002/02/0022/)

DETECTION OF THE CORE STRUCTURE OF TRANSCRIPTION FACTOR BINDING SITES

[1] Pozdnyakov M.A., [2] Vityaev E.E., [1] Ananko E.A., [1] Busygina T.V., [1] Ignatieva E.V., [1] Proskura A.L., [1] Podkolodnaya O.A., [3] Podkolodny N.L., [1] Merkulova T.I., [1] Kolchanov N.A.

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[2] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[3] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, 630090, Russia
e-mail: mike@bionet.nsc.ru
*Corresponding author

Key words: *transcription, recognition, binding sites, transcription factors, sequence alignment*

Resume

Motivation: Computer analysis and recognition of transcription factor binding sites (TFBS) are important for a better understanding of eukaryotic gene expression regulation. However, current methods for TFBS recognition are relatively imprecise. To recognize TFBS more precisely, specific contextual features of TFBS sequences should be taken in account, such as the number and localization of conservative regions, the length of spacers separating the conservative regions, etc.

Results: A method for determining the number and localization of the conservative regions (cores) within TFBS was developed. The core regions of binding sites of the factors COUP, SF-1, and STAT agree well with the published data.

Introduction

Regulation of eukaryotic gene transcription involves many proteins, including transcription factors (TF; Kolchanov et al., 2002). TF bind to certain DNA regions called transcription factor binding sites (TFBS). A computer-assisted annotation of TFBS in long genome sequences will undoubtedly advance our understanding of the mechanisms underlying genome functioning. At present, there are numerous methods for TFBS recognition based on the construction of consensuses and weight matrices (Quandt et al., 1995; Werner, 2000; Prestridge, 2000; Poznyakov et al., 2001), neural networks (Demeler, Zhou, 1991; Pedersen, Engelbrecht, 1995; Ogura et al., 1997), etc. However, methods for precise TFBS recognition have not been developed yet. There are several reasons underlying this problem: (i) a vast number of transcription factors (at present, several thousands of TF are known, see Brivanlou, Darnell, 2002); (ii) a great diversity in the mechanisms of DNA–protein interactions between TFBS and transcription factors; (iii) complexity of TFBS organization and diversity in their contextual, physicochemical, and conformational properties (Ponomarenko et al., 1999); and (iv) a context surrounding TFBS, which is specific for various regulatory regions (promoters, enhancers, silencers, locus control regions, etc.).

Specific of TFBS is that various types of these sites might have a certain number of core (conservative) regions separated by variable regions (spacers). The development of methods for TFBS recognition has been severely hampered by the lack of *a priori* information on the number, orientation, and localization of the core regions within the sites. Therefore, of great importance is the development of computer tools for analysis of contextual properties of TFBS, which will provide us with information on the number and localization of the core regions within the sites. Consequently, allowance for such a complex site structure will undoubtedly facilitate the development of more precise methods for TFBS recognition.

Materials and Methods

SF-1, COUP, and STAT binding site nucleotide sequences were retrieved from the EMBL database, basing on the information on the localization of the sites stored in the TRRD database (Kolchanov et al., 2002). The samples of the TFBS of SF-1, COUP, and STAT included 39, 33, and 29 sites, respectively. The site sequences had a length of 30 bp each.

The method developed, which allows the determination of the number, localization, and boundaries of conservative and variable regions in the sequences sampled, involves the following steps:

1) Multiple alignment of the TFBS sequences (Fig. 1a);

2) Construction of a frequency matrix using the alignment results (Fig. 1b);

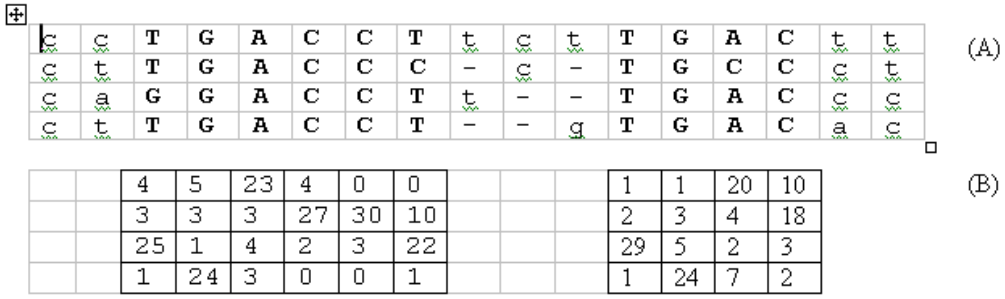
3) ***Analysis of the frequency matrix constructed*** (Fig. 1*b*)**.** For every matrix position, the value of χ^2 was determined, which characterizes nonuniformity of absolute base frequencies at the current position. Let W_a, W_c, W_t, and W_g be absolute base frequencies at a specified matrix position and $W = W_a + W_c + W_t + W_g$ be the sum of base frequencies in a

matrix column. Then, E_x, the expected number of a base X ($X = a,c,t,g$), equals $E_x = W \times Q_x$, where Q_x is the frequency of the X base in the human genome. The nonuniformity of the absolute base frequencies in the matrix column is

$$\chi^2 = \sum_{x=a,c,t,g} \frac{(W_x - E_x)^2}{E_x} \quad (1)$$

4) ***Construction of χ^2 profile for frequency matrix positions.*** Each of the maximums on the profile that exceeds the value of χ^2 averaged over all the matrix positions is considered as a core region of the TFBS of a particular type. Figure 1 shows the frequency matrices corresponding to two core regions of the COUP factor binding site.

5) ***Localization of the core regions in each site of the initial sample.*** A set of the core regions detected for each site of the initial sample was determined by a pairwise alignment of a site sequence with a set of the frequency matrices corresponding to these core regions. The method of aligning a site with a set of matrices is illustrated in Fig. 2.



c	c	T	G	A	C	C	T	t	c	t	T	G	A	C	t	t	(A)
c	t	T	G	A	C	C	C	–	c	–	T	G	C	C	c	t	
c	a	G	G	A	C	C	T	t	–	–	T	G	A	C	c	c	
c	t	T	G	A	C	C	T	–	–	g	T	G	A	C	a	c	

4	5	23	4	0	0				1	1	20	10
3	3	3	27	30	10				2	3	4	18
25	1	4	2	3	22				29	5	2	3
1	24	3	0	0	1				1	24	7	2

(B)

Fig. 1. (A) Fragment of multiple alignment of the COUP factor binding sites. The core regions of the sites are capitalized and bold. (B) Frequency matrices for the core regions of the sites constructed by multiple alignment.

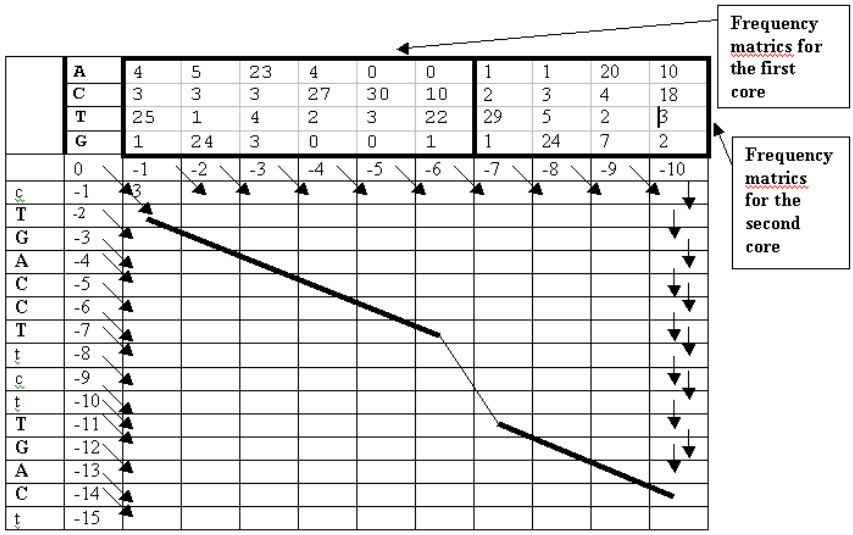


Fig. 2. Scheme of the alignment of the COUP factor binding site sequence (site sequence is shown as a vertical on the left) with the set of frequency matrices constructed for core regions. There are two frequency matrices on the scheme. The first four rows of the second column have symbols A, C, T, and G, specifying the base corresponding to each matrix row. Short arrows illustrate the steps for the dynamic matrix filling (not shown for every dynamic matrix element). Bold lines indicate the correspondence between the core regions of the site sequence with the frequency matrices.

Results and Discussion

Using the approach described, we studied the binding sites of the factors SF-1, COUP, and STAT. Figure 3 shows χ^2 profiles of the weight matrices for the SF-1, COUP, and STAT binding sites.

The SF-1 sites have one core region and the COUP and STAT sites have two core regions each, which agrees well with the published data. Using frequency matrices, the corresponding consensus sites were constructed that agree with those detected previously: TCAAGGTCA for SF-1 (Busygina et al., 2000), TGACCT(n)$_{2–4}$ TGACCT for COUP (Ladias et al., 1992), and TTNNNNNAA for STAT (Seidel et al., 1995) (see Fig. 3).

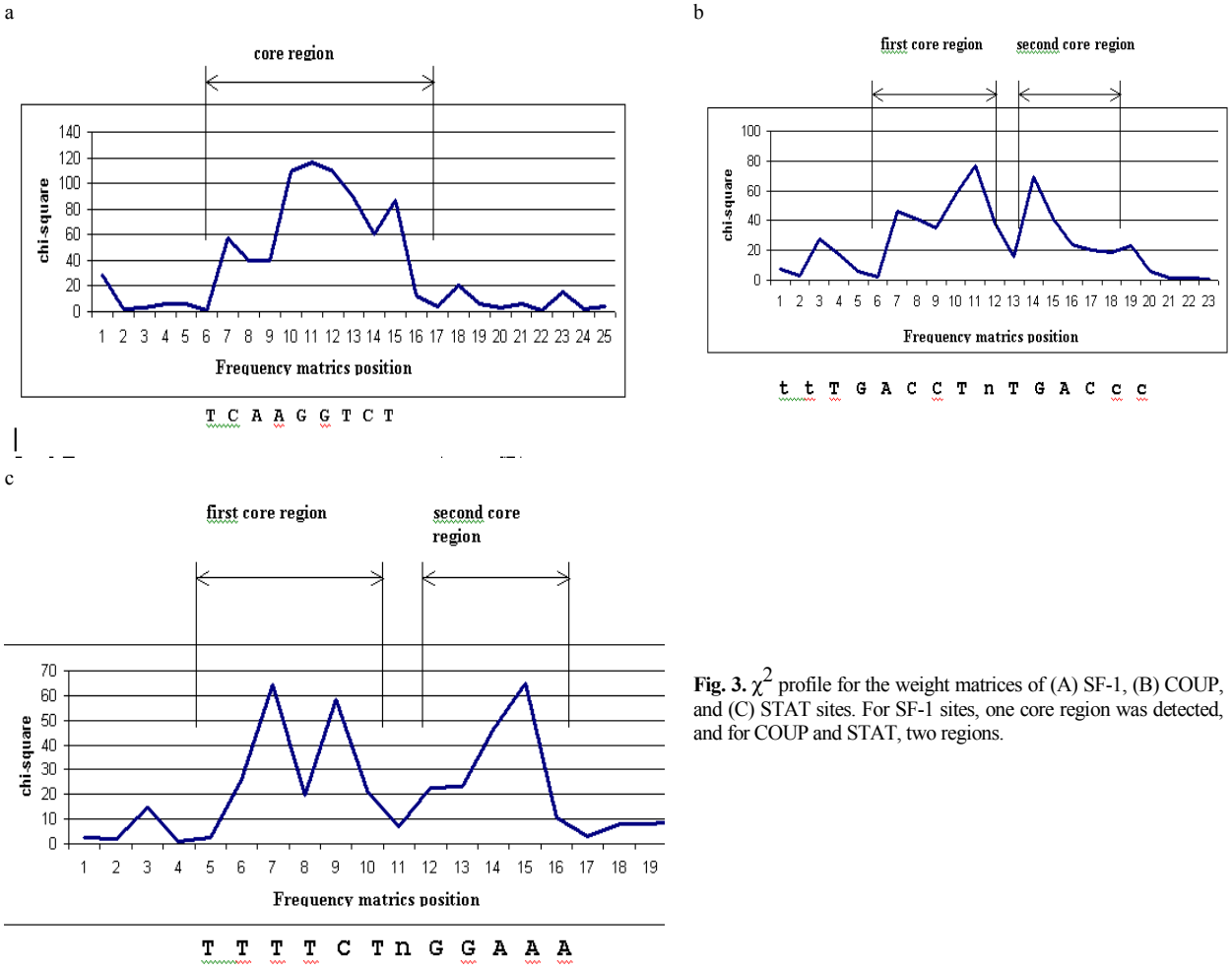


Fig. 3. χ^2 profile for the weight matrices of (A) SF-1, (B) COUP, and (C) STAT sites. For SF-1 sites, one core region was detected, and for COUP and STAT, two regions.

Acknowledgements

The authors thank A.V.Osadchuk for his assistance in selection of SF-1 samples.

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 07-90337, 00-02-07-90355, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (subgrant № 535228 CFDA 81.049).

References

1. Brivanlou A.H., Darnell Jr., J.E. (2002). Signal transduction and the control of gene expression. Science. 295(5556): 813–818.
2. Busygina T.V., Ignatieva E.V., Osadchuk A.V. (2000). Steroidogenesis-controlling gene transcription regulation: representation in TRRD database. In Proc. Second Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2000), (ICG, Novosibirsk, August 7-11) 1:41–44.
3. Demeler B., Zhou G. (1991) Neural network optimization for *E. coli* promoter prediction. Nucl. Acids Res. 19:1593–1599.
4. 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. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312–317.
5. Ladias J.A.A., Hadzopoulou-Cladaras M., Kardassis D., Cardot P., Cheng J., Zannis V., Claradas C. (1992). Transcriptional regulation of human apolipoprotein genes APOB, APOCIII, and APOAII by members of the steroid hormone receptor superfamily HNF-4, ARP-1, EAR-2, and EAR-3. J. Biol. Chem. 267:15849–15860.

6. Ogura H., Agata H., Xie M., Odaka T., Furutani H. (1997). A study of learning splice sites of DNA sequence by neural networks. Comput. Biol. Med. 27:67–75.
7. Pedersen A.G., Engelbrecht J. (1995). Investigations of *Escherichia coli* promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional startpoint. Intelligent Systems Mol. Biol. 3:292–299.
8. Ponomarenko J.V, Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., Kolchanov N.A. (1999). Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics. 15:654–668.
9. Pozdnyakov M.A., Vityaev E.E., Ananko E.A., Ignatieva E.V., Podkolodnaya O.A., Podkolodny N.L., Lavryushev S.V., Kolchanov N.A. (2001). Comparative analysis of methods recognizing potential transcription factor binding sites. Mol. Biol. 35:961–969.
10. Prestridge D.S. (2000). Computer software for eukaryotic promoter analysis. Methods Mol. Biol. 130:265–295.
11. Quandt K., Frech K., Karas H., Wingender E., Werner T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucl. Acids Res. 23:4878-4884.
12. Seidel H.M., Milocco L.H. Lamb P., Darnell J.E., Stein R.B., Rosen J. (1995). Spacing of palindromic half sites as a determinant of selective STAT (signal transducers and activators of transcription) DNA binding and transcriptional activity. Proc. Natl Acad. Sci. USA. 92:3041–3045.
13. Werner T. (2000). Computer-assisted analysis of transcription control regions. Matinspector and other programs. Methods Mol. Biol. 132:337–349.

EXPRESSION OF LIPID METABOLISM GENES: DESCRIPTION IN TRRD DATABASE AND COMPUTER-ASSISTED ANALYSIS

Proscura A.L., Levitsky V.G., Oshchepkov D.Yu., Pozdnyakov M.A., * Ignatieva E.V.

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

Key words: *database, LM-TRRD, transcription regulation, lipid metabolism, COUP-TF, SRE, HNF4, PPRE*

Resume

Motivation: Systematization, generalization, and analysis of knowledge on transcription regulation of genes of the lipid metabolism are the actual tasks, because disruptions of lipid metabolism cause a seria of severe human diseases. Accumulation of data on nucleotide sequences of transcription factor binding sites (TFBS) that regulate the functioning of genes of lipid metabolism, as well as computer-assisted analysis of these TFBS, may be a source of valuable novel information about regularities in organization of these regulatory gene regions.

Results: We have developed a novel release of the database LM-TRRD on regulation of transcription of genes of lipid metabolism. On the basis of LM-TRRD, we have compiled the samples of TFBS of four types: COUP-TF, SRE, HNF4, and PPRE. We have analyzed the distribution of sites in regulatory regions of genes and studied the nucleosome potential of the regions of DNA in vicinities of the sites. We have designed the software programs for recognition of the sites of two types, COUP-TF and PPRE.

Availability: LM-TRRD is available by the Internet via the address http://www.bionet.nsc.ru/trrd/

Introduction

LM-TRRD (Lipid Metabolism – Transcription Regulatory Regions Database) is one of thematical sections of TRRD database that accumulates information about transcription regulation of the genes controlling lipid metabolism (Ignatieva et al., 1997). Previously, we have analyzed the content of transcription factors that regulate transcription of genes in this group. As shown, transcription factors SREBP, HNF4, COUP-TF, and PPAR play an important role in regulation of genes of this group (Ignatieva et al., 2000). In the course of developing the knowledge base on lipid metabolism, we accumulate information about TFBS and develop the methods for their recognition. In the present work, we study two characteristics of TFBS, namely, their distribution relatively transcription start and distribution of nucleosome potential within the neighboring DNA regions. By using samples of binding sites of two transcription factors, COUP-TF and PPAR, we have developed the software programs for recognition of relevant sites by the method SITECON (Oshchepkov et al, this issue). These programs were applied for recognizing potential binding sites of these two transcription factors in 5'-flanking regions of genes from LM-TRRD.

Methods and Algorithms

Data accumulation in LM-TRRD was made by experts-biologists by means of annotating the experimental publications in the format of TRRD database (Kolchanov et al., 2002).

Samples of TFBS were compiled in accordance with the data stored in TRRD database (release 6.01), by using the program TRRD-Sample developed by authors. The main criterion of including a TFBS into the sample was the strict experimental verification of interaction between transcription factor of the given type and DNA (i.e., experiments on DNase I footprint with purified protein, electrophoretic mobility shift assay (EMSA) with purified protein, EMSA with nuclear cell extract and with antibodies to a protein, as well as experiments on co-transfection of transcription factor studied together with artificial mutagenesis).

The nucleosome potential profile of values within DNA regions surrounding TFBS was examined by the method based on discriminate analysis and on accounting dinucleotide frequencies in the local regions of nucleosomal sites (Levitsky et al., 2001) (http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/).

Recognition of TFBSs (COUP-TF and PPRE) in the 5'-flanking regions of lipid metabolism genes was performed by the SITECON method (Oshchepkov et al., 2002). For developing the method aimed at PPRE recognition, we have used the sample consisting of 14 sites, of 40 bp in length. Recognition was made at 86% conformational similarity level. False positive estimate equals to 20% (the control was provided by the jack-knife method, with subsequent removing of 7% of sequences). False positive estimate was 1.6 x 10-3 (1/600). For developing the method recognizing COUP-TF site, we have used the sample consisting of 33 sequences, of 27 bp in length, under the level of conformational similarity of 76%.

The false positive estimate equaled to 20% (the control was made by jack-knife method, removing 10% of sequences). False negative estimate equaled to 2.6 x 10-3 (1/372). For predicting potential sites, COUP-TF and PPRE, we have used the sample of regulatory regions (-600/-1 relatively transcription start) of 17-teen genes from the section of TRRD database, LM-TRRD. Nucleotide sequences were extracted from the EMBL databank by the program TRRD-Pars developed by us.

Implementation and Results

Informational content of LM-TRRD

In the current release of TRRD database, the section LM-TRRD is supplemented with plenty of novel information. Since 2000, the informational content has been increased: genes, 1.5 fold; regulatory regions, 1.7 fold; TFBS, 1.5 fold; bibliography, 1.8 fold (Figure 1). The LM-TRRD database is a constituent part of the GeneExpress system (Kolchanov et al., 2002) and it is available by the address http://www.bionet.nsc.ru/trrd/.

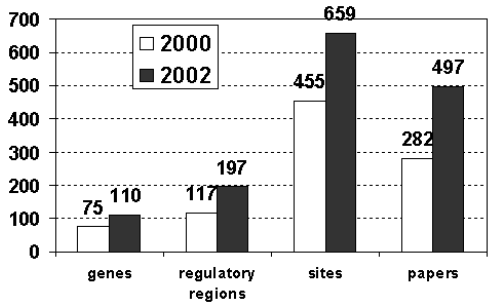


Fig. 1. Informational content of the LM-TRRD database during 2000 - 2002

Samples of TFBS

Based on accumulated in TRRD database data, we have compiled the samples of TFBS of four types: COUP-TF, SRE, HNF4, and PPRE. Each sample is a set of fragments of regulatory regions of genes, with the lengths of 300 and 120 nucleotides, respectively, including experimentally characterized site in a central position. Informational content of the samples is illustrated in Table.

Table. Samples of TFBS used in this work.

Sample name	Site length, in bp	Site name/transcription factor name	Number of sequences
coup_120	120	COUP-TF / COUP-TF (Chicken Ovalbumin Upstream Promoter Transcription Factor)	33
coup_300	300		28
sre_120	120	SRE* / SREBP (Sterol Regulatory Element Binding Protein)	27
sre_300	300		23
hnf4_120	120	HNF4 / HNF4 (Hepatic Nuclear Factor 4)	35
hnf4_300	300		31
ppre_120	120	PPRE** / PPAR (Peroxisome Proliferator-Activated Receptor)	16
ppre_300	300		14

* Sterol Regulatory Element; ** Peroxisome Proliferator Responsive Element.

GENE DISCOVERY COMPUTER SYSTEM FOR ANALYSIS OF REGULATORY REGIONS

[1] Vityaev E.E., [2] Pozdnyakov M.A., [2*] Orlov Yu.L., [2] Vishnevsky O.V., [2] Podkolodny N.L., [2] Kolchanov N.A.

[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: orlov@bionet.nsc.ru
[*]Corresponding author

Key words: *machine learning, knowledge discovery, data mining, bioinformatics, eukaryotic promoter recognition, transcription factor binding sites*

Resume

Motivation: Analysis of gene regulatory regions requires an integrated study of heterogeneous information encoded in DNA sequences at the level of DNA binding sites of regulatory proteins and at the level of protein–protein interaction of transcription factors.

Results: Gene Discovery, a computer system, was developed to trace patterns in the contextual organization of DNA sequences and, basing on the results obtained, predict transcription factor binding sites (TFBS) and regulatory regions. The system allows a user to (1) find local patterns of the contextual organization of TFBS, (2) detect mechanisms of distribution of potential TFBS, and (3) collect the data on hierarchical organization of promoter regions and recognize these regions using the data obtained. Samples of promoters of several functional systems stored in the TRRD database were analyzed, and the patterns of their contextual organization were detected, which yield a more precise recognition.

Availability: The software application for Windows is available upon request from the authors.

Introduction

Importance of Hierarchical Analysis of Promoters

Investigation in the structure of regulatory regions is promising for a better understanding of regulatory mechanisms of eukaryotic gene transcription. Each regulatory region contains binding sites of specific transcription factors (TFBS) (Nikolov and Burley, 1997). One gene can have many promoters that determine the formation of various protein products or show different levels of specific functional activity. However, context signals of eukaryotic promoters, which are very important for their functions, are typically weak and localized to a wide region (Goodrich et al., 1996).

Prediction of regulatory and, first and foremost, promoter regions requires integration of heterogeneous information encoded in DNA sequences at the level of DNA–protein binding and interaction of transcription factors.

Recognition of Sites and Promoters by their Contexts

A diverse structure of gene promoter regions complicates the development of software applications for promoter recognition. Despite a large number of methods for recognizing RNA polymerase II promoters in eukaryotic genomes (Pedersen et al., 1999; Kondrakhin et al., 1995), a more precise recognition remains a serious problem.

However, the recognition of TFBS is a particular problem, which can be solved by using weight matrices, expert rules, or multiple alignments (Pozdnyakov et al., 2001). A protein binding site is not a linear simple structure within a nucleotide sequence; in contrast, it has conservative (core) and variable regions, which are assessed by the corresponding positions of the weight matrix. To define the boundaries of variable regions, it is necessary to have a larger number of sequences in the sample.

The *Gene Discovery* system detects the patterns of the module organization of linear sequences.

Methods and Algorithms

Gene Discovery (Fig. 1) is an implementation of the *Discovery* system used for analysis of promoter regions (Vityaev et al., 2001) and for other life sciences fields (Kovalerchuk, Vityaev, 2000).

Contrast samples of nucleotide sequences belonging to the following two classes are inputted into the system:

1) Functional class—promoters or sequences of sites binding to flanks (positive sample) and

2) Alternative class—DNA sequences, which do not perform this function (negative sample).

A package of software applications is available that searches for contextual signals in the sequences of these two classes (Fig. 1). The signal means a short interval of the analyzed sequence (word). The signal can be

Contextual (a short oligonucleotide word, a functional site, etc.);

Conformational (a DNA region with specific conformational or physicochemical properties, such as low-melting DNA regions, strongly bent DNA structures, etc.); and

Structural (for example, Z-DNA, a hairpin RNA secondary structure, etc.).

These signals can be detected using *a priori* knowledge based on experimental information retrieved from specialized databases (Kolchanov et al., 2002). In addition, the signals can be determined using the programs for detection of regulatory signals, which employ consensuses, weight matrices, or profiles of conformational DNA properties and trained on other data (Babenko et al., 1999). The combinations of such signals located linearly within a DNA sequence are the desired complex signal patterns of the contextual structure of the sequence.

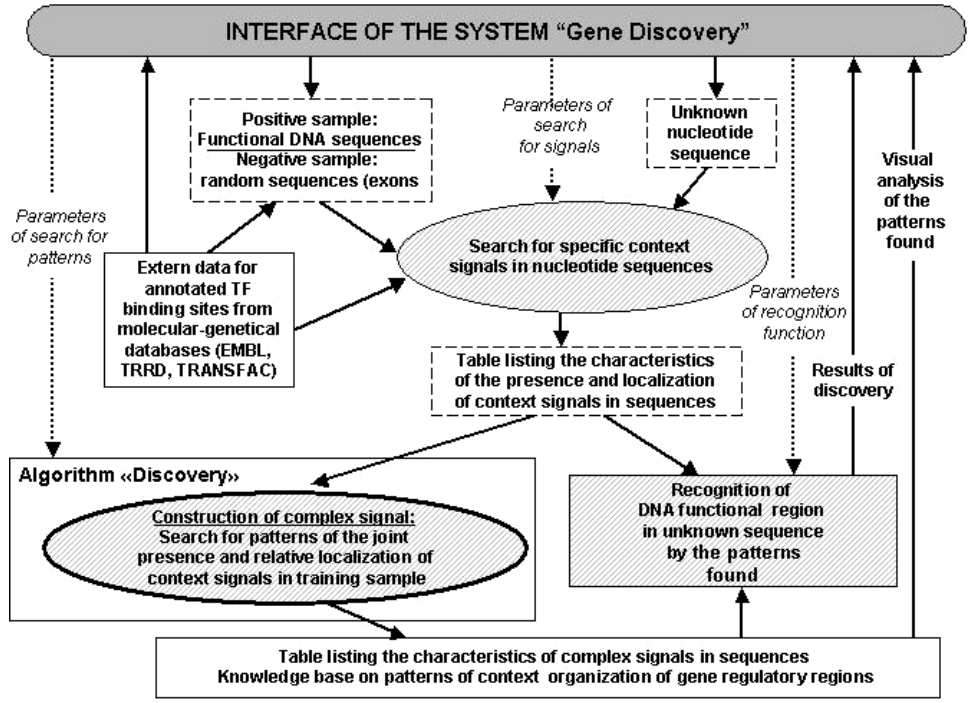


Fig. 1. Scheme of *Gene Discovery* system.

Vityaev et al. (2001; 2002) studied complex signals consisting of only one type of signals—imperfect oligonucleotides. Oligonucleotide signals S_1, S_2, etc., were written as words of a length of 8 bases using a generalized 15 single letter–based IUPAC code: {A, T, G, C, R = G/A, Y = T/C, M = A/C, K = T/G, W = A/T, S = G/C, B = T/C/G, V = A/G/C, H = A/T/C, D = A/T/G, and N = A/T/G/C}. The complex signal was searched for in a form of $(S_1,S_2,..S_k)$, where $k>1$ is the total number of elements in the complex signal (oligonucleotide order with respect to the transcription start was taken into account).

We present an extended version of the program, which allows construction of the complex signals as (1) functional sites consisting of short conservative oligonucleotides down to mononucleotides; (2) groups of binding sites (regulatory regions) consisting of degenerate oligonucleotides of a length of 5–10 bp; and (3) hierarchical regulatory regions consisting of complex signals detected by the program at the first stage.

Complex signals are constructed as patterns, which are correlated with a sequence and might be related to TFBS and other functional sites, for example, human splice sites (Fig. 2).

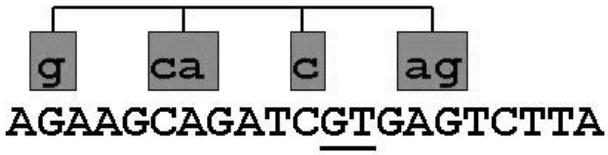


Fig. 2. An example of the pattern g<c<a<c<a<g used for a splice site sequence. Selected GT is conservative region in donor splice sites.

Signals can be present accidentally in a negative sample of sequences and be absent in some sequences of the functional class. The total number of patterns might be large enough. In this case, the totality of sequences cannot be covered by patterns even at the stage of training. Such a "nonstrict" approach to constructing complex signals is due to a large variety

of potential TFBS in the regulatory regions and a problem of construction of a training sample of jointly regulated genes. An analogue of these patterns can be a method that "implements" a variant of the site located within the sequence analyzed (Kondrakhin et al., 1995).

Recognition of control data implemented in the program offers the following options how to use the patterns determined: (1) selection and application of the most statistically significant patterns (that displayed a better accuracy in the training sample) and (2) estimation of the homology to the sequences displaying the same patterns. The recognition method based on oligonucleotide signals is described in (Vishnevsky, Vityaev, 2001).

Implementation and Results

Gene promoter sequences of functional systems retrieved from the TRRD database were analyzed (Kolchanov et al., 2002), including erythroid-specific gene promoters, promoter regions for cell cycle regulatory genes, promoters of genes controlling lipid metabolism, and promoters of genes expressed in muscle. Promoter sequences had a length of 120 bp (from –100 to +20 bp with respect to the transcription start site). As estimated, homology between the promoters did not exceed 60% for each pair. Complex signals were built as groups of potential binding sites.

From the viewpoint of molecular biology, the patterns detected are unique complex signals that are of primary importance for the proper function of a promoter (Vityaev et al., 2001).

Discussion

The main result of this work is the development of a new method for analysis and recognition of regulatory regions. Using an optimized process of construction of complex signals, a user can select best signals for a rather short time, about several hours of working on a Pentium-II PC, without laborious sorting out the totality of variants. This method can be applied to a wide range of problems, for example, analysis of functional sites of all types with identification of specific signatures of DNA bases. In this case, the general homology among the sequences might be rather weak, and only some core regions located in different parts of the sequence and separated from each other by spacers of various lengths might show conservative characteristics.

This approach was successfully tested by analysis of regulatory regions and binding sites of proteins functioning as dimers and oligomers. The hierarchical recognition takes into account the module structure of regulatory regions (Klingenhoff et al., 1999). Undoubtedly, detection and allowance for complex signals will increase the precision and accuracy of recognition of specific promoter groups.

Acknowledgements

The authors thank E.A.Ananko and E.V.Ignatieva for their assistance with data retrieval and helpful discussions. The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 00-07-90337, 02-07-90355, 00-04-49229, and 02-01-00939); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Babenko V.N. et al. (1999) Investigating extended regulatory regions of genomic DNA sequences. Bioinformatics. 15:644–653.
2. Goodrich J.A. et al. (1996) Contacts in context: promoter specificity and macromolecular interactions in transcription. Cell. 84:825–830.
3. Klingenhoff A. et al. (1999) Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. Bioinformatics. 15:180–186.
4. Kolchanov N.A. et al. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312–317.
5. Kovalerchuk B., Vityaev E. (2000) Data Mining in Finance: Advances in Relational and Hybrid Methods. Kluwer Acad. Publ.
6. Kondrakhin Y.V. et al. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. Comput. Appl. Biosci. 11:477–488.
7. Nikolov D.B., Burley S.K. (1997) RNA Polymerase II transcription initiation: a structural view. Proc. Natl Acad. Sci. USA. 94:15–22.
8. Pedersen A.G. et al. (1999) The biology of eukaryotic promoter prediction—a review. Comput. Chem. 23:191–207.
9. Poznyakov M.A. et al. (2001) Comparative analysis of methods recognizing potential transcription factor binding sites. Mol. Biol. (Mosk.). 35:961–978.
10. Vishnevsky O.V., Vityaev E.E. (2001) Analysis and recognition of promoters of the erythroid-specific genes on the basis of degenerated oligonucleotide motifs. Mol. Biol. (Mosk.). 35:979–986.
11. Vityaev E.E. et al. (2001) Computer system "Gene Discovery" for regularities retrieving in eukaryotic regulatory sequences organization. Mol. Biol. (Mosk.). 35:952–960.
12. Vityaev E.E. et al. (2002) Computer system "Gene Discovery" for promoter structure analysis. In Silico Biol. 2:0024 http://www.bioinfo.de/isb/2002/02/0024/

AUTHOR INDEX

KEY WORDS