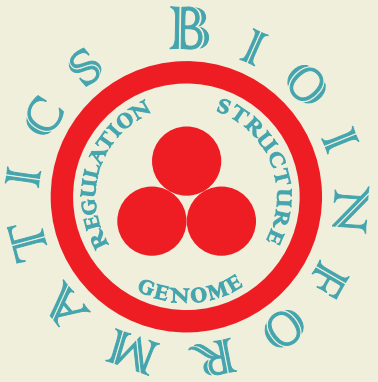
PROCEEDINGS OF THE FIFTH INTERNATIONAL CONFERENCE ON BIOINFORMATICS OF GENOME REGULATION AND STRUCTURE



VOLUME 1

BGRS' 2006
NOVOSIBIRSK, RUSSIA
JULY 16 - 22, 2006

RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS

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

Edited by
N. Kolchanov, R. Hofestädt

Volume 1

BGRS'2006
Novosibirsk, Russia
July 16–22, 2006

Novosibirsk
2006

INTERNATIONAL PROGRAM COMMITTEE

Nikolay Kolchanov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
(Chairman of the Conference)
Ralf Hofestadt University of Bielefeld, Germany
(Co-Chairman of the Conference)
Dagmara Furman Institute of Cytology and Genetics SB RAS, Novosibirsk,
(Conference Scientific Secretary)
Dmitry Afonnikov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Mikhail Gelfand GosNIIGenetika, Moscow, Russia
Vadim Govorun Institute of Physicochemical Medicine, RAMS, Moscow, Russia
Reinhart Heinrich Humboldt University Berlin, Berlin, Germany
Charlie Hodgman Multidisciplinary Centre for Integrative Biology,
School of Biosciences, University of Nottingham, UK
Alexey Kochetov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Eugene Koonin National Center for Biotechnology Information,
National Library of Medicine, National Institutes of Health, Bethesda, USA
Vassily Lyubetsky Institute for Informational Transmission Problems RAS,
Moscow, Russia
Luciano Milanesi National Research Council – Institute of Biomedical Technology, Italy
Viatcheslav Mordvinov Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Yuriy Orlov Genome Institute of Singapore,
Laboratory of Information & Mathematical Sciences, Singapore
Igor Rogozin Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
Cenk Sahinalp Computing Science, Simon Fraser University, Burnaby, Canada
Maria Samsonova St.Petersburg State Polytechnic University, St.Petersburg, Russia
Akinori Sarai Kyushu Institute of Technology (KIT), Iizuka, Japan
Konstantin Skryabin Centre "Bioengineering" RAS, Moscow, Russia
Rustem Tchuraev Institute of Biology, Ufa Sci. Centre RAS, Ufa, Russia
Denis Thieffry ESIL-GBMA, Universite de la Mediterranee, Marseille, France
Jennifer Trelewicz IBM Almaden Research Center, San Jose, California, USA
Edgar Wingender UKG, University of Goettingen, Goettingen, Germany
Lev Zhivotovsky Institute of General Genetics RAS, Moscow, Russia
Jagath C. Rajapakse School of Computer Engineering,
Nanyang Technological University, Singapore

LOCAL ORGANIZING COMMITTEE

Sergey Lavryushev Institute of Cytology and Genetics SB RAS, Novosibirsk
(Chairperson)
Ekaterina Denisova Institute of Cytology and Genetics SB RAS, Novosibirsk
Andrey Kharkevich Institute of Cytology and Genetics SB RAS, Novosibirsk
Galina Kiseleva Institute of Cytology and Genetics SB RAS, Novosibirsk
Anna Onchukova Institute of Cytology and Genetics SB RAS, Novosibirsk
Yuri Orlov Institute of Cytology and Genetics, Novosibirsk
Natalia Sournina Institute of Cytology and Genetics SB RAS, Novosibirsk

ISBN 5-7692-0846-5 (V.1)
ISBN 5-7692-0845-7

Our sponsors

Organizers

Laboratory of Theoretical Genetics
of the Institute of Cytology and Genetics, SB RAS

Institute of Cytology and Genetics, SB RAS

Siberian Branch of the Russian Academy of Sciences

Vavilov Society of Geneticists and Breeders

Scientific Council on Bioinformatics,
Siberian Branch of the Russian Academy of Sciences

The Chair of Informational Biology
of the Department of Natural Sciences
of Novosibirsk State University

Grants

INTAS Grant

Russian Foundation for Basic Research

Contents

Introduction

Three volumes of *Proceedings of the Fifth International Conference on Bioinformatics of Genome Regulation and Structure–BGRS'2006* (Akademgorodok, Novosibirsk, Russia, July 16–22, 2006) comprise about 200 peer-reviewed publications on the topical problems in bioinformatics of genome regulation and structure. Biology now is among the most dynamically developing scientific disciplines. The main factor of this progress is an unprecedented, both in the rate and volume, accumulation of new facts due to advent of novel state-of-the-art experimental technologies. The post-genome era in biology brought about a sharp up-scaling of the research in the fields of genomics, transcriptomics, and proteomics. We are the witnesses how new directions of experimental and computer molecular biology emerge and successfully advance, including sequencing and analysis of megagenomes of bacterial communities, regulation of gene expression by short RNAs, microarray analysis technique, construction of proteomic portraits of cells and tissues, metabolomics, high-throughput genotyping of human populations for biomedical purposes, and many others. However, the synthesis of these directions is developing to a lesser degree, while it is a primary need for creation of an orderly theory of development, function, and evolution of the living systems—systems biology (gene interaction, gene network functioning, signal transduction pathways, networks of protein–protein interactions, modeling of ontogenesis, molecular phylogeny, the theory of evolution, etc.). The reasons underlying this gap lie not only in the objective complexity of the living systems, but also in the specialization in various fields of biology, which is ever increasing with accumulation of new data and development of new methods. The holistic vision of the research object is disappearing. The goal of this Conference, similar to the preceding Conferences—BGRS'1998, BGRS'2000, BGRS'2002, and BGRS'2004, which were held in Novosibirsk in 1998, 2000, 2002, and 2004—is, first and foremost, to provide the possibility for a wide exchange of opinions for various experts in *in silico* biology and researchers involved in experimental studies who use computer methods in their work or have interest in applied or theoretical aspects of bioinformatics. BGRS'2006 provides a general forum for disseminating and facilitating the latest developments in bioinformatics in molecular biology. BGRS'2006 is a multidisciplinary conference. The scope covered by the Conference comprises (i) the issues of development of advanced methods for computational and theoretical analysis of structure–function genome organization, proteomics, transcriptomics microarray analysis, etc.; (ii) application of these methods in theoretical (various aspects of evolutionary biology) and applied (search for promising application points in biotechnology and medicine) fields; and (iii) the issues related to general informational support of biological research and education (creation and computer support of databases, retrieval systems, ontologies, etc.). Thus, the final goal of this Conference may be defined as a half the battle for the new synthesis in Biology, which is a long-standing need, via the dialogue

between the experts in particular fields of biology. This is the reason why BGRS'2006, along with the traditional sections (computational structural and functional genomics and transcriptomics, computational structural and functional proteomics, comparative and evolutionary genomics and proteomics, and bioinformatics and education), includes an essentially expanded section on ***computational systems biology,*** which contains the presentations on modeling of molecular genetic systems and processes in bacterial and multicellular organisms and modeling of morphogenesis. Moreover, as compared to the previous conferences, the presentations related to evolution and phylogeny are plentiful. Numerous interdisciplinary studies into various taxa performed by the methods of molecular phylogeny, computer genomics, proteomics, cytogenetics, etc., as well as comparison of these results with the data obtained by classical methods of evolutionary morphology, paleontology, and various directions of ecology revealed the basic differences between the rates and modes of evolution at different hierarchical levels of biological organization (genes, genomes, karyotypes, organisms, populations, and biocenoses). Thus, the actual evolutionary process cannot be reduced to the evolution on one of the listed levels and is, speaking in images, an interference pattern, which is the more complex, the more interacting blocks and hierarchical levels constitute a biological system and the more intricate are their interrelations. Deciphering of this interference pattern is one of the challenges for the biology of the XXI century, which is answerable only by the joint efforts of bioinformatics and experimental sciences. If BGRS'2006 succeeds in contributing to this to any degree, the organizers will reckon their goal fulfilled.

Among the main goals of BGRS is improvement in the quality of education in all its aspects. That is why the success and international acknowledgement of the preceding conferences and the 2005 BGRS Summer School "Evolution, Systems Biology and High Performance Computing Bioinformatics" has encouraged launching the 2006 BGRS Summer School "Evolution, Systems Biology and High Performance Computing Bioinformatics". This School being the co-event of the conference will precede BGRS'2006. This event will attract next generation of scientists to bioinformatics. The scientific scope of the school will include issues of the development and application of advanced methods of computational and theoretical analysis for structure-function genome organization, proteomics, evolutionary and systems biology. We hope that the School of Young Scientists will become a good BGRS tradition.

BGRS'2006 is organized by the Laboratory of Theoretical Genetics with the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, (Novosibirsk, Russia). The organizational sponsors of the Conference are the Institute of Cytology and Genetics and the Siberian Branch of the Russian Academy of Sciences. The financial sponsor is the Russian Foundation for Basic Research. The School of Young Scientists "Evolution, Systems Biology and High Performance Computing Bioinformatics" is sponsored by the Russian Foundation for Basic Research and INTAS. The organizational support for the School is provided by the Chair of the Informational Biology, Faculty of the Natural Sciences of the Novosibirsk State University and the Council of Young Scientists of the Institute of Cytology and Genetics, SB RAS.

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

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

BGRS 2006

PART 1. COMPUTATIONAL STRUCTURAL AND FUNCTIONAL GENOMICS AND TRANSCRIPTOMICS

NEW WAY TO OBTAIN A REGULATORY MOTIF REPRESENTATION DUE TO MOTIF ABUNDANCE LEVEL

***Abnizova I.*[*1], *Walter K.*[1], *te Boekhorst R.*[2], *Gilks W.R.*[1]**
[1] Biostatistics Unit MRC, Institute of Public Health, Robinson Way, CB2 2SR, Cambridge, UK;
[2] Computer Science Department, University of Hertfordshire, College Lane, AL10 92BA, Hatfield Campus, UK
[*] Corresponding author: e-mail: irina.abnizova@mrc-bsu.cam.ac.uk

Key words: statistical approach, transcription factor binding sites, motifs, association

SUMMARY

Motivation: An important step in understanding of gene regulation is the recognition of gene expression regulatory elements and regions. Experimental procedures for this are slow and expensive. We present a novel statistical approach to show the association of experimentally verified regulatory elements with over-represented motifs within regulatory regions, together with a way to recognize these regulatory regions and make a consensus motif description using available online tools.

Results: In our method, we exploit the fundamental property of regulatory regions: the abundance of over-represented transcription factor binding motifs. The method provides a way to find these over-represented motifs, in the form of exceptionally large lists of similar words, and construct their consensus descriptions. We rank the motifs due to their abundance level. The association of experimentally confirmed binding sites and predicted motifs allows the method to be potentially used as complementary tool for motif discovery.

Availability: The source code is available at the http://www.mrc-bsu.cam.ac.uk/BSUsite/AboutUs/People/irina.xml.

INTRODUCTION

One of the great challenges in bioinformatics is to understand the varied and complex mechanisms that regulate gene expression. We focus on one important step of this problem, the statistical characterisation of regulatory regions, and establish an association between putative regulatory elements and over-represented motifs within them.

Regulatory regions, comparatively short sequences (several hundred to several thousand base pairs, depending on the species) upstream or downstream of the transcription start site often play a major role in the regulation of gene expression. The study of regulatory DNA is more difficult than that of coding sequences (Wasserman *et al*., 2000; Dermitzakis, Clark, 2002). There are no well known properties in regulatory DNA analogous to open reading frames and non-uniform codon usage in coding sequences. This makes it difficult to define the consensus and location of functional regulatory elements, at specific sites within regulatory regions, recognized by regulatory proteins (transcription factors), which act upon binding as transcriptional repressors or activators, controlling the rate of transcription. Revealing the statistical properties typical of regulatory regions and regulatory elements may improve our understanding of their evolutionary and functional constraints.

A number of computational algorithms designed to search for functional regulatory elements using evolutionary comparisons, whole-genome data, and putative co-regulated genes have been successfully demonstrated in recent years. It should not be assumed, however, that all functional sequences are conserved, and all non-functional have diverged. It is also problematic to correctly define a set of co-regulated genes. We suggest a statistical approach which may be used as a complementary tool for motif discovery. We have tested our approach to find on the annotated regulatory regions of approximately 19 genes for which experimentally validated transcription factor binding sites (TFBSs) are readily available. We developed the software allowing a description of motifs ranked due to their abundance.

We describe here a content based method to characterise regulatory regions and to assess an association between regulatory elements and over-represented motifs within them. We assume that the abundance of regulatory motifs within regulatory regions leaves a distinct "signature" in nucleotide composition, and that it is possible to capture this "signature" statistically. More specifically, we hypothesize that it takes the form of an over-representation of "similar words" (which are not simple repeats). This over-representation should show up as outliers in the right tail of the distribution of similar word lists of variable length. We identify such outliers, present these word lists as a consensus strings, using the well-known WebLogo tool (http://weblogo.berkeley.edu/), and show their association with known TFBS.

MATERIALS AND METHODS

We assess the association of TFBS and over-represented motifs in regulatory DNA in three steps:

1. We construct the distribution of similar words in a stretch of genomic DNA. We infer its putative function using our "fluffy-tail-test" (Abnizova *et al*., 2005).
2. If the sequence passes the test, we identify a number of significantly over-represented motifs in the form of families of similar words, which we call the *maximal similar word lists* (MSWLs).
3. Score the presence of experimentally confirmed TFBS in these lists with Z-scores, and assess its statistical significance.

Then we rank the most high scored lists, and submit them into WebLogo to obtain their consensus description. Note that we call two words of the same length k-similar, if they have k mismatches. Thus, for example, two words "aacctg" and "cacctg" are 1-similar. Generally speaking, one can run the algorithm for any word length, m (m less than the sequence length), and number of mismatches, k. In the work (Abnizova *et al*., 2005) we used m = 3,5,7, 9,12 with corresponding k = 0,1,2,3 to infer a putative regulatory function of a given DNA stretch.

We tested the association of confirmed instances of TFBS with our MSWLs on the set of experimentally verified *Drosopila melanogaster* regulatory regions provided by Papatsenko *et al*. (2002). The set consists of 19 regulatory regions from early developmental genes, with annotated locations of TFBS. Each regulatory sequence is from 700 up to 1600 bp long, containing from 10 to 25 annotated TFBS.

RESULTS AND DISCUSSION

In the test data sequences 84 % (16 out of 19) were found to have stronger association of TFBS and MSWL than by chance: $Z \geq 2$. The results for all annotated 19 regulatory regions are summarised in Table 1. Note that the sequences without significant association are actually did not pass the "fluffy-tail" test of being regulatory DNA, see last three rows in Table 1.

Table 1. The association of TFBS and MSWL for annotated regulatory regions

Name of regulatory region	Z score	N_{real}	μ	s
AbdominA	3.7	89	30.09	15.6
Hairy str7	2.1	299	146.4	69.7
Hairy str5	2.45	49	14.5	14.02
Even-sk.srt37	4.2	232	98.7	31.3
Even-sk.srt2	2.01	45	19.5	12.4
Engrailed intron	2.2	44	8.2	25.7
Tailles	2.15	51	32.7	8.47
Ult pbx	4.4	152	62.9	20.1
Runt5	2.17	107	18.4	66.7
Spalt early	2.05	187	135.0	27.2
kruppel	2.7	186	66.2	33.3
Hairy6	4.62	1358	543.4	182.1
Orthodent	5.36	1720	679.2	195.9
ftp	2.08	48	30.7	8.29
Even-sk.srt46	4.06	146	45.2	24.8
gooseberry	2.02	88	28.7	30.1
buttonhead	-0.46	54	68.1	30.1
Ult bre	-0.44	13	15.3	5.1
ftz	-0.01	43	43.3	21.0

Note. Key to the included genes: AbdominA –Abdominal Anterior enhancer, hairy strj –hairy stripe (with number j) enhancer, Even-sk.srtij – even-skipped stripe ij enhancer, ult pbx – ultrabithorax proximal regulatory region, ftp – fushi-tarazu proximal enhancer, gooseberry – gooseberry enhancer, ftz – fushi-tarazu zebra enhancer, ulr bre – ultrabithorax enhancer, orthodent – orthodentical enhancer, buttonhead – buttonhead cis element. N_{real} stands for maximal list size in original sequence, μ and s are mean and standard deviation of maximal list sizes in randomised sequences.

We submit these significant MSWL into WebLogo (Crooks *et al*., 2004) to obtain the motifs. One can see an example of such a list for MSWL associated with the experimentally verified hunchback TFBS. As a result, we obtain the 'portrait' of our most abundant motif within the sequence for fushi-tarazu proximal enhancer region as following:

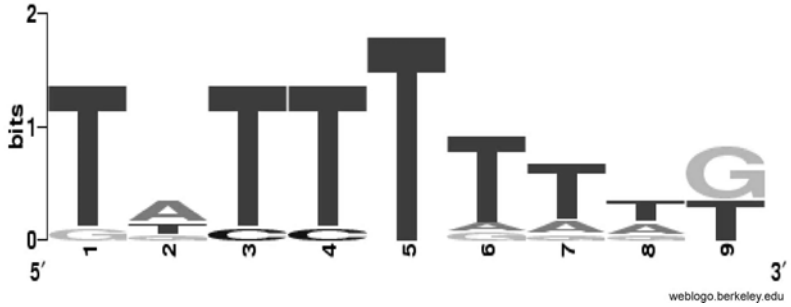


Figure 1. The web-logo description of most abundant list of similar words within experimentally verified and annotated *fushi-tarazu* proximal enhancer region. It was found to be associated with ttk annotated TFBS.

Compare the description of the abundant motif above in the Fig. 1 with the real experimentally verified *tramtrack* TFBS instances and their description in Fig. 2. This MSWL is associated with *tramtrack* binding sites within the *fushi-tarazu* proximal enhancer region, and they are reasonably consistent (CAGGAC consensus parts in both Fig. 1 and 2):

Figure 2. Instances of experimentally verified ttk TFBS within the *fushi-tarazu* proximal enhancer.

The main message of our method is the ability to pick up true TFBS within significant MSWL, and to reconstruct their "generalised" consensus from predicted MSWL. However, our method in its current form is not yet a motif discovery tool. In our fluffy tail test, the main feature distinguishing regulatory DNA from other genomic DNA is the presence of exceptionally large families of similar motifs, producing fluffy right tails in the distribution of similar list length. These motifs differ from simple tandem repeats due to their spatial arrangement (see Abnizova *et al*., 2005). From the point of view of motif discovery our MSWLs would contain many false positive instances of TFBS. However, these "false positive" instances constitute a strong signal and might have an important biological role of attracting TFs. For this reason, in our future work we would like to filter MSWL by utilising other motif discovery methods. Our results indicate a significant statistical association of over-represented motifs with experimentally confirmed TFBS in confirmed cis-regulatory modules. This association suggests our methodology might be of value, in combination with other strategies, for motif discovery.

REFERENCES

Abnizova I., Walter K., te Boekhorst R., Gilks W.R. (2005) Some statistical properties of regulatory DNA sequences, and their use in predicting regulatory regions in the *Drosophila* genome: the fluffy-tail test. *BMC Bioinformatics*, **6**, 109.

Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. (2004) WebLogo: a sequence logo generator. Genome Res., **14**, 1188–1190.

Dermitzakis E., Clark A. (2002) Differential selection after duplication in mammalian developmental genes. *Mol. Biol. Evol.*, **18**, 557–562.

Papatsenko D.A., Makeev V.J., Lifanov A.P., Regnier M., Nazina A.G., Desplan C. (2002) Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers. *Genome Res.*, **12**, 470–481.

Wasserman W., Palumbo M., Thompson W., Fickett J., Lawrence C. (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet*., **26**, 225–228.

A KNOWLEDGE AND DATA BASED HYBRID APPROACH TO GENE CLUSTERING

Abhishek K.[1*], Karnick H.[1], Mitra P.[2]
[1] Indian Institute of Technology Kanpur, Kanpur, India; [2] Indian Institute of Technology Kharagpur, Kharagpur, India
[*]Corresponding author: e-mail: kumabhi@iitk.ac.in

Key words: gene ontology, gene expression, clustering, minimum spanning tree, cluster score, cluster quality

SUMMARY

Motivation: Traditional gene clustering algorithms focus only on the raw expression data for clustering genes whereas valuable information about genes is available in the form of GO trees. We aim to use this information along with expression data to produce better gene clusters.

Results: We propose an algorithm that produces good quality cohesive clusters and does not require the a priori specification of the number of clusters. The proposed algorithm comprehensively outperforms the k-means and random clustering algorithms on two yeast cell data sets.

Availability: Available on request.

INTRODUCTION

Microarray technology produces expression data for thousands of genes under different conditions. Interpretation of this huge volume of observed data requires the clustering of genes that have correlated expression profiles. Clustering can help identify genes that may have common functions or those that may be part of common regulatory networks. Traditional clustering algorithms have focused on the raw gene expression data for performing this task (Brazma, Vilo, 2000). However, valuable biological knowledge in the form of the Gene Ontology (GO) can provide useful inputs for producing meaningful clusters. The GO represents terms in a directed acyclic graph (DAG), covering three taxonomies namely molecular function, biological process, and cellular component. For example, the gene product cytochrome can be described as follows: molecular function terms: oxidoreductase activity; biological process terms: oxidative phosphorylation, induction of cell death; component terms: mitochondrial matrix, mitochondrial inner membrane. The DAG consists of terms represented as nodes connected by relationship edges. The ontology annotates gene products with different terms across the graph. In this work we adopt a hybrid approach towards gene clustering. We use knowledge available in the form of GO together with gene expression data to perform clustering.

ALGORITHM

Distance measure between GO nodes. We use the GO process ontology, to find the distance between GO nodes. Each GO node is annotated with a list of genes; some nodes can be un-annotated. Since GO is a DAG a GO node can have multiple parents and multiple children. As a particular node can have multiple parents the GO DAG is transformed to get a **directed tree structure.** The **level** of a node is defined as the number of nodes between the node and the root node. The procedure for converting the DAG to a directed tree is as follows:

Do {

Visit every node x in the DAG

If node x has k>1 parent then:

- Create k nodes and make each newly created node a single child of a distinct parent of x.
- Replicate annotations of x across the k newly created nodes.
- Make each child of x a child of each newly created node.
- Remove node x.

} **Until** (No change in DAG)

We define a **weight function** f:{1,2, …, N} → R where N is the maximum level of any node in the obtained directed tree and R is the real line such that f is a decreasing function i.e. f(i) > f(j) for all i < j.

The distance between nodes i and j in the directed tree is defined to be the weight of the level of the **Least Common Ancestor** (LCA) of the nodes i and j. The arguments in support of choosing this as a distance measure and the proof that it is in fact a distance metric can be found in Lee et al. 2004.

Calculation of gene distance matrix from the directed GO tree. We calculate the distance between all gene pairs which are annotated in the directed GO tree. We first prepare an n*m table T where n is the total number of genes and m is the number of nodes in the GO tree. The rows represent the gene and the columns represent the attributes or GO nodes. We treat each GO node as an attribute. As a single gene may be annotated to multiple nodes we check for the list of annotations of a single gene and put a 1 in every attribute column with which the particular gene is annotated. So we get a binary table, which is used for calculating gene distance. We define diff(i, j) to be the number of differing entries in corresponding columns of rows i and j. The gene distance between genes g_i and g_j $GOdist(g_i,g_j)$ is defined to be: $GOdist(g_i,g_j) = \frac{1}{diff(i,j)} * \sum_{1<=\alpha<=n} \sum_{1<=\beta<=m} (T_{i\alpha} - T_{i\beta})^2 * d(\alpha,\beta)$; $d(\alpha,\beta)$ is the GO distance between GO nodes α and β. We get a gene distance matrix G of the order n*n where n is the number of genes.

Definitions:

- Average GO distance $GOavg(C)$ corresponding to a gene cluster C = g_i, 1 < = i < = k, $GOavg(C) = \sum_{1<=i,j<=k} GOdist(g_i, g_j) \Big/ C_2^k$; $GOdist(g_i, g_j)$ is the GO distance measure between g_i and g_j based on the gene distance matrix G.
- Scatter of a cluster $Scatter(C)$ C = g_i,1 < = i < = k, where g_i's are the members of the cluster C $Scatter(C) = \sum_{1<=i<=k} (x_i - \mu)(x_i - \mu)^T$; where x_i is the expression vector of gene g_i and μ is the average expression vector for the cluster C. $(\mathbf{x_i}-\boldsymbol{\mu})^T$ is the transpose of the vector $\mathbf{x_i} - \boldsymbol{\mu}$.
- The objective function for partitioning the MST for a given number of clusters k F(k):

$$F(k) = \frac{1}{k} * \sum_{1<=i<=k} (GOavg(C_i) + Scatter(C_i))$$

- $Score_i$ is the minimum value of the objective function F (i) obtained for a given number of clusters i and $Cluster_i$ is the optimal cluster set corresponding to it.

Iterative Clustering Algorithm. The sequence of steps followed by the algorithm is as follows:

1. Calculation of expression data distance matrix: Use Euclidean measure to calculate gene expression data matrix E of order n*n.

2. Scale both gene distance matrix G and expression data matrix E to the same range and combine them to get net distance D.

3. Make a fully connected graph with genes as nodes with edge weight between node i and node j equal to the distance between gene i and gene j obtained from the net distance matrix D. Find the minimum spanning tree of this fully connected graph. To cluster, partition this Minimum Spanning Tree (MST) into k sub trees where k is the number of desired clusters (Xu *et al*., 2002).

4. The iterative algorithm is as follows:

INPUT: MST obtained from step 3.

a. Initialize k to 1.

b. while k < MAXCLUSTERS{

c. Perform a random k-partitioning of the Minimum Spanning Tree (MST) by removing k-1 edges. Then perform the following operation until the process converges. For each pair of adjacent clusters, go through all the edges in the merged cluster of the two to find the edge to cut, this globally optimizes the 2-partitioning of the merged cluster measured by objective function F(k).

d. $Score_k$ = F(k), save the optimal cluster set obtained in above step as $Cluster_k$

e. Increment k.

}

f. Search the list of scores to find the minimum element $Score_{min}$ and output the cluster set $Cluster_{min}$ corresponding to it.

RESULTS AND CONCLUSION

We used the data set of the Yeast cell cycle in which activity was measured at 18 time points. We used two subsets each consisting of 500 genes for testing the algorithm. Cluster validation was done using figure of merit score (FOM). We compared the proposed algorithm with k-means, random, and our algorithm without using GO distances. Fig. of Merit score (FOM) is defined as in (Yeung *et al*., 2001) suppose e is the left out condition of the m experimental conditions present in the data set, let there be k clusters $C_1, C_2, \ldots C_k$, and let $E_{g,e}$ be the expression level of g under condition e. Let $\mu_{Ci}(e)$ be the average expression level in condition e for genes of cluster C_i.

$$FOM(e,k) = \sqrt{\frac{1}{n} * \sum_{1<i<=k} \sum_{x \in C_i} (E_{x,e} - \mu_{C_i}(e))^2} \text{ and } FOM(k) = \sum_{1<=e<=m} FOM(e,k).$$

We use $FOM(k)$ to assess the quality of clusters obtained from different algorithms for a given number of clusters k. It is basically a leave one out approach, where clustering is performed using all but one of the experimental conditions in the data set. The left out condition is used to assess the predictive power of the clustering algorithm. The FOM score represents scatter from the actual value at test condition; thus lower the FOM score higher is the predictive power of the algorithm and better is the quality of clusters obtained.

The performance graph for the four algorithms on the two data sets is shown in Fig. 1*a*, *b*. The following can be observed from the plots of Fig. 1*a*: (a) The proposed algorithm outperforms k-means and the random algorithm for greater than 10 clusters. (b) The proposed algorithm without GO distances outperforms k-means after 20 clusters. (c)

The proposed algorithm without GO distances outperforms random algorithm right from the beginning. (d) The proposed algorithm outperforms the one without GO distances right from the beginning. Performance on the second data set (Fig. 1*b*) reveals the following: (*a*) The proposed algorithm with and without GO distances outperforms k-means and the random algorithm right from the beginning. (*b*) The proposed algorithm outperforms the one without GO distances right from the beginning. We conclude from these observations that the proposed algorithm outperforms k-means and the random algorithm and the usage of GO distances improves cluster quality.

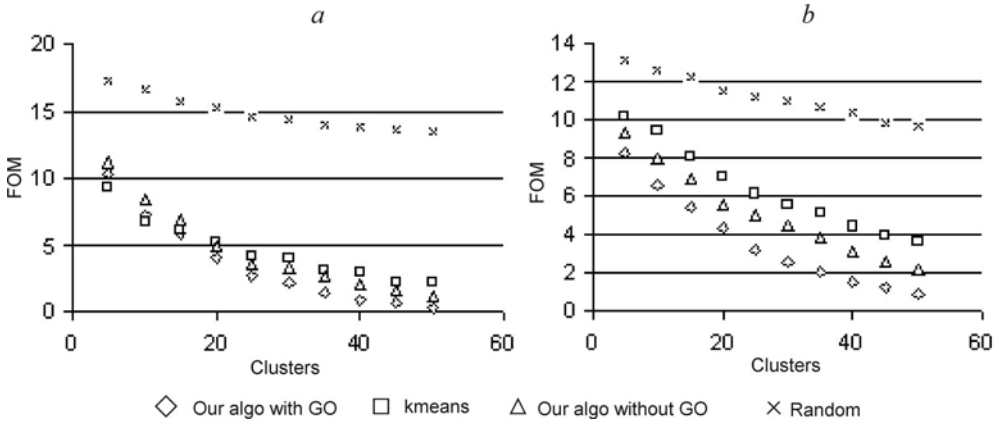


Figure 1. Performance analysis on dataset1 (*a*); performance analysis on dataset2 (*b*).

REFERENCES

Brazma A., Vilo J. (2000) Gene expression data analysis. *FEBS Lett*, **480**, 17–24.

Lee S.G. *et al*. (2004) A graph theoretic modeling on GO spaces for biological interpretation of gene clusters. *Bioinformatics*, **20**(3), 381–388.

Xu Y. *et al*. (2002) Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning tree. *Bioinformatics*, **18**(4), 536–545.

Yeung K.Y. *et al*. (2001) Validating clustering for gene expression data. *Bioinformatics*, **17**(4), 309–318.

www.geneontology.org/GO.docs.html

PREDICTION OF INTERFERON-INDUCIBLE GENES IN HUMAN GENOME

Ananko E.A.*, Kondrakhin Yu.V., Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: eananko@bionet.nsc.ru

Key words: recognition of transcription factor binding sites, interferon-stimulated genes, genome annotation

SUMMARY

Motivation: Application of the methods of computer-assisted genome annotation coupled with large-scale experimental studies may be helpful in determining possible functions of numerous unstudied genes. The search for interferon-inducible genes is of particular interest. As known, interferons modulate the work of the immune system: they exert antiviral, antibacterial, and antitumoral effect. Although the system of interferons is being actively studied during several dozens years, the mechanisms of its functioning are still not clear in many respects.

Results: By using the methods developed for recognition of interferon-inducible genes, an analysis of DNA sequences of more than 2000 genes within the length limits from -1000 to +1000 bp relatively transcription start was performed. We have detected 78 genes that could be interferon-inducible with high probability and could participate in supporting some interferon's functions.

Availability: The list of predicted interferon-inducible human genes obtained in the course of the work considered is available at http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/iig-trrd/ISG_predicted.html.

INTRODUCTION

Interferons (IFNs)[1] are classified into two types: IFNs type I, or virus-inducible acid-resistant interferons (e.g., leukocyte IFN-α, fibroblast IFN-β, IFN-δ, IFN-ε, IFN-κ, IFN-ω, IFN-τ, and IFN-z) and IFN type II, or immune acid-liable IFN-γ. Type I interferons are mainly support antiviral state of the organism (Pestka *et al*., 2004), whereas IFN-γ makes larger impact in providing antibacterial and antiparasitic responses (Decker *et al*., 2002). Also, IFN-γ was shown to participate in development of autoimmune states (Baccala *et al.*, 2005).

By studying signal transduction pathways of interferon system, it was established that IFNs type I cause activation of ISGF3 transcription factor, whereas IFN-γ – activation of the STAT1 homodimer (Platanias, 2005). Type I interferons activate also Akt serine-threonine kinase and p38/MAP-kinase cascades, as well as the signal transduction pathways leading to activation of NF-κB and p53 transcription factors: all these factors participate in the antiviral immune response and tumor suppression (Pestka *et al.*, 2004;

[1] The abbreviations used are: bp, base pair; IFN, interferon; ISG, Interferon-Stimulated Genes; ISGF3, Interferon-Stimulated Gene Factor 3; IRF, Interferon Regulatory Factor; STAT, Signal Transducer and Activator of Transcription.

Platanias, 2005). Interaction of these and some other transcription factors with transcription factor binding sites in regulatory regions of interferon-stimulated genes (ISG) mediate significant increase of gene transcription.

METHODS AND ALGORITHMS

For recognition of transcription factor binding sites, we have used an additive recognition function with application of statistic simulation approach (Kondrakhin *et al.*, 2006).

For selecting individual sites for recognition of ISG, we have applied a statistical test that compares two binomial values, that is, mainly those sites were selected, relative occurrence frequency of which within the interferon-inducible genes (F1) was statistically increased in comparison to that of the genes entering the control sample of genes extracted from the EPD database (F2). To select a pair of sites, we have used the standard statistical test χ^2. For majority of selected sites and pairs of sites, p-value, was < 0.001.

The score at each position of arbitrarily chosen sequence SEQ0 was calculated as follows. First, we have selected **m** objects from the training sample (i.e., sites and pairs of sites), T_1, T_2, …, T_m. Then for the i-th position of the sequence SEQ0, we calculate **m** of weights w_1,w_2, …, w_m by the formula:

w_i = {1, if the i-th object, T_i, is not found at respective positions of the sequence SEQ0;
F1/F2, in case the i-th object, T_i, is found at respective positions of the sequence SEQ0}.

Then we calculate the score by multiplying the weights w_1, w_2, …, w_m , so that

$$\text{SCORE} = w_1 * w_2 * \ldots * w_m . \quad (1)$$

Then we calculate the score for all positions of the sequence SEQ0 and select the position with the maximal score.

Notably, the more is the number of the objects selected at relevant positions of the sequence SEQ0, the higher is the score. In other words, SCORE is a function measuring similarity between the sequence SEQ0 studied and the training sample, out of which the objects T_1, T_2, …, T_m were extracted.

For calculation of SCORE for each method (see results), the same multiplicative function (1) was applied, but for each method its own set of pre-selected objects (i.e., sites and pairs of sites), T_1, T_2, …, T_m, was used.

RESULTS

By using three methods of ISG recognition (Kondrakhin *et al.*, 2006), we have studied 1664 human genes, within the regions from -1000 to +1000 bp relatively transcription start site, annotated in the EPD database. In order to minimize type II error, we have ordered very stringent threshold limits for all these three methods applied simultaneously. The threshold value of recognition function for the method 0 (induction by any IFN) equals to 0.4. The values of the other two recognition functions should also exceed the threshold level equaling to 0.4 for the method 1 (induction by type I IFNs) and 0.3 for the method 2 (induction by type II IFN).

The verification of recognition methods was accomplished by using the sample of ISG that were identified by microarray data (sample M0 for the positive control). In Table 1, the results of recognition of IFN-inducible genes in different samples of genes are given. In addition to the training sample ISG-TRRD that was compiled on the basis of the TRRD database (Kolchanov *et al.*, 2002), and the sample M0 for the positive control, we have also tested the sample compiled on the basis of EPD database (1664 human genes). As the negative control, we have analyzed two samples containing very small percentage of ISG,

i.e., genes regulated by glucocorticoids (GR-TRRD) and genes of lipid metabolism (LM-TRRD). Recognition was performed under the same conditions for all the samples: the sequences from -1000 to +1000 relatively transcription start site were analyzed.

In total, among 1664 human genes extracted from the EPD database, we have found 78 genes that potentially response to stimulation by interferons (Table 1). Four genes out of 78 were previously included into the training sample. For 60 genes detected, the stimulation by interferons was not reported yet. In addition, for 13 genes, experimental evidence was obtained demonstrating that transcription is enhanced under the action of interferons by means of RNA microarray data. In 28 genes found, the regions of maximal sensitivity to interferon induction were located in promoter region (from -200 to+50 bp relatively transcription start).

Table 1. Recognition of interferon-inducible genes among different samples under the threshold limitations equaling to 0.4 for the method 0 and method 1, and 0.3 for the method 2

Sample	Sample size (total number of sequences)	Genes recognized (total number)	Genes recognized (%)
ISG-TRRD	72	17	23.6
M0	1005	156	15.5
EPD	1664	78	4.7
GR-TRRD	70	1	1.7
LM-TRRD	58	0	0

DISCUSSION

The potential ISGs found may be classified into several functional groups with respect to biological activities of interferons as the genes involved into (i) immune and inflammatory response, (ii) regulation of cell proliferation and differentiation, and (iii) antitumoral effect (see the complete list of genes at http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/iig-trrd/ISG_predicted.html). For 21 genes, it was difficult to relate their activity with biological function of interferons, so they are considered as possible over-estimation.

Due to our estimates, by taking into account possible over-estimation (in total, 21 genes out of 78 recognized, or 1 % of the sample), human genome carries about 3000 ISG. This value does not contradict to microarray data. For example, only in the primary culture of monocytes isolated from peripheral blood of patients diseased by hepatitis C, at least two-fold induction of 1012 genes was registered under the action of IFN-α during 6 hours after simulation (Ji *et al.*, 2003), whereas in IFN-γ-stimulated macrophages, 632 genes were induced (Ehrt *et al.*, 2001). In hepatocarcinoma cell line HepG2, out of 14 112 genes considered, more than 400 genes were induced by two-fold by IFN-α and 405 genes were induced by IFN-γ (Xiong *et al.*, 2003).

Simultaneous application of computer-assisted methods for recognition of genes simulated by various IFNs enables to reveal in mammalian genome with high accuracy ISG that are involved in interferon system functioning.

ACKNOWLEDGEMENTS

The authors are grateful to E.V. Ignatieva for kindly providing the sample of lipid metabolism genes and to G.V. Orlova for translating the manuscript from Russian into English. The work was supported by the Russian Government (Contracts Nos 02.434.11.3004, 02.467.11.1005) and Siberian Branch of the Russian Academy of

Sciences (the project "Computational simulation and experimental design of gene networks" and integration project No. 115).

REFERENCES

Baccala R., Kono D.H., Theofilopoulos A.N. (2005) Interferons as pathogenic effectors in autoimmunity. *Immunol Rev.*, **204**, 9–26.

Decker T., Stockinger S., Karaghiosoff M., Muller M., Kovarik P. (2002) IFNs and STATs in innate immunity to microorganisms. *J. Clin. Invest*, **109**, 1271–1277.

Ehrt S., Schnappinger D., Bekiranov S., Drenkow J., Shi S., Gingeras T.R., Gaasterland T., Schoolnik G., Nathan C. (2001) Reprogramming of the macrophage transcriptome in response to interferon-gamma and *Mycobacterium tuberculosis*: signaling roles of nitric oxide synthase-2 and phagocyte oxidase. *J. Exp. Med.*, **194**, 1123–1140.

Ji X., Cheung R., Cooper S., Li Q., Greenberg H.B., He X.S. (2003) Interferon alfa regulated gene expression in patients initiating interferon treatment for chronic hepatitis C. *Hepatology*, **37**, 610–621.

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.

Kondrakhin Yu.V., Ananko E.A., Merkulova T.I. (2006) Methods for recognition of interferon-inducible sites, promoters, and enhancers. *This issue.*

Pestka S., Krause C.D., Walter M.R. (2004) Interferons, interferon-like cytokines, and their receptors. *Immunol. Rev.*, **202**, 8–32.

Platanias L.C. (2005) Mechanisms of type-I- and type-II-interferon-mediated signalling. *Nat. Rev. Immunol.*, **5**, 375–386.

Xiong W., Wang X., Liu X.Y., Xiang L., Zheng L.J., Liu J.X., Yuan Z.H. (2003) Analysis of gene expression in hepatitis B virus transfected cell line induced by interferon. *Sheng Wu Hua Xue Yu Sheng Wu Wu Li Xue Bao (Shanghai),* **35**, 1053–1060.

A DATABASE DESIGNED FOR THE POLYMORPHISMS OF THE HUMAN CCR2 GENE

Apasyeva N.V.[1], Yudin N.S.[1*], Ignatieva E.V.[1], Voevoda M.I.[1, 2], Romashenko A.G.[1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Institute of Internal Medicine, SB RAMS, Novosibirsk, Russia
[*] Corresponding author: e-mail: yudin@bionet.nsc.ru

Key words: database, human CCR2 gene, polymorphism, disease, trait, population, allele frequency

SUMMERY

Motivation: Abundant information about all the currently known human genomic polymorphic markers is stored in the databases, whose sophisticated structure makes difficult efficient search of the required information. The development of a specialized secondary database with the information presented more compactly can substantially facilitate the user's work.

Results: We developed a specialized database that contains information about the polymorphic markers of the CCR2 gene and neighboring DNA regions, population frequencies of certain polymorphisms and SNP associated diseases and traits. The database can be useful for extracting *in silico* the polymorphisms of the CCR2 gene that have causal effect on the pathogenesis of diseases associated with immune system responses.

Availability: The database is available on request from the authors.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) are currently the most informative markers for the genes that cause common complex diseases. SNP are more abundant (1 SNP per 100–1,000 bp), their detection is cheaper and less labor consuming than that of the other genomic polymorphic markers. Information about the SNPs and other polymorphic markers of the human genome is stored in the well known free available databases, including dbSNP, HGVbase, OMIM, and others. However, the bulky universal archives have complicated structures, this poses obstacles to search of the needed information about genetic markers and associations with diseases. Besides, these archives usually do not contain the data about the traits and diseases, because manual annotation of the continually expanding scientific information is required. Assembly of the data for polymorphisms in more specialized databases would allow to store them in the more compact and accessible format. Thus, the user's job to tracking polymorphisms would be facilitated.

The CCR2 gene is of major interest with reference to certain widespread threatening diseases (AIDS, cancer, diabetes) (Le *et al.*, 2004). We have previously demonstrated that substitution of valine by isoleucine at position 64 of the protein sequence (V64I) is associated with myocardial infarction (MI) (Voevoda *et al.*, 2002). Support for this association subsequently came from two independent teams. However, the detection of the association does not yet mean that this particular polymorphism is the cause of disease predisposition. It may be located on a chromosome nearby another truly disease-causative polymorphism. The database for the human CCR2 polymorphisms is required for search of the causative polymorphism at the predisposition locus to MI. It is hoped that the database we newly created would be a helpful tool to researchers dealing with the CCR2 gene.

METHODS AND ALGORITHMS

Search in the Internet resourses was done by using the National Center for Biotechnology Information Service (http://www.ncbi.nlm.nih.gov/). The database was created as tables on the MS Excel format. Hyperlinks to the respective URL were added manually when required. Annotated abstracts and full article texts were the main sources for filling up the database.

IMPLEMENTATION AND RESULTS

We have developed a specialized database that contains information about polymorphic markers (predominantly SNPs) in the CCR2 gene and its neighboring DNA regions, their population frequencies and also about the trait and diseases associated with these polymorphisms. The database consists of 4 interrelated tables. Table "GENE" contains the general information about the gene: its complete and short names, references to the cards of the gene in the databases EntrezGene, GeneCards, EMBL/GenBank, NCBI, references to the protein card in the SwissProt databases. Table "POLYMORPHISMS" contains the following data about polymorphisms: identification number (rs#) in the dbSNP database, nucleotide position in the chromosomal contig, positions of the substituted amino acids in the protein, validation status. The information includes also the polymorphism effect on the gene expression level (if available), links to cards in the NCBI, UCSC and SwissProt databases, references to the published literature data; nucleotide sequences from the dbSNP database that flank the polymorphic site are additionally provided. The "“DISEASES" table lists the names of the diseases, SNP-diseases association (yes or no), ethnic group, sex and age of the subjects in the examined sample. The "POPULATIONS" table includes the country and region, the name of the examined population, the frequencies of the minor allele and of its genotypes. The third and fourth tables contain hyperlinks to the original publications and to their abstracts in the PubMed database, and currently contain information about SNP CCR2-64I only.

The compiled database contains information about 41 polymorphisms. Besides 36 SNPs, the database provides information about 4 single nucleotide and 1 dinucleotide deletions. 4 polymorphisms are located in the promoter region, 21 are in the first intron (Fig. 1, positions 42757–46055), 2 are in the second intron (positions 47047–48254) 6 are in the coding parts (positions 46106–47046; 48255–48438) of exons (of these, 3 are nonsynonymous), 4 are in the 3'UTR (positions 48439–49505), and 4 are in the 3′-flanking distant region of the gene. The classification of the polymorphisms was based on the structure of the longest mRNA (isoform A). The validation status of 6 polymorphisms is "unknown"; the existence of others is experimentally supported. There are 66 units that describe disease associations and 131 information units for the population frequencies. We intend to further annotate and improve the database.

DISCUSSION

The CCR2 gene has 3 exons and covers about 7 kb on human chromosome 3p21 (Fig. 1). There are two known alternative mRNA isoforms, A and B. Both have identical 5′-ends composed of exon 1 (positions 42728-42756) and the 5′-part of exon 2 (positions 46056–47046), but they differ by the mRNA regions encoding their carboxyl end and 3′UTR. The two isoforms code the functional receptors differing by subcellular localization. The 41 polymorphisms in the database are unevenly distributed according to gene nucleotide sequence. SNPs are densest in the first intron, rare SNPs occur in the second. The average polymorphism density (1 SNP per 200 bp) agrees with their average density in the human genome. However, because we presented the total data for the numerous samples that

differed by ethnic group, sex, age and other features, the SNP list will increase with time. Evidently, the information about SNP associations with certain diseases is inconsistent. We revealed 5 publications that examined the CCR2-64I association with MI (Gonzalez *et al.*, 2001; Voevoda *et al.*, 2002; Ortlepp *et al.*, 2003; Petrkova *et al.*, 2003; Bjarnadottir *et al.*, 2005). The association was proven in 3 of the 5 (Voevoda *et al.*, 2002; Ortlepp *et al.*, 2003; Petrkova *et al.*, 2003). Probable environmental (diet, lifestyle etc) and/or genetic factors that may abolish the association between CCR2-64I and MI are yet to be found. It is of interest that in the popular OMIM database the CCR2 card does not mention association with MI. Regrettably, the OMIM is advisable so far as a preliminary introduction to the problem. The information about the population frequencies of the CCR2-64I allele would allow comparing the frequency of this polymorphism with morbidity and mortality of cardiovascular diseases among a population under study.

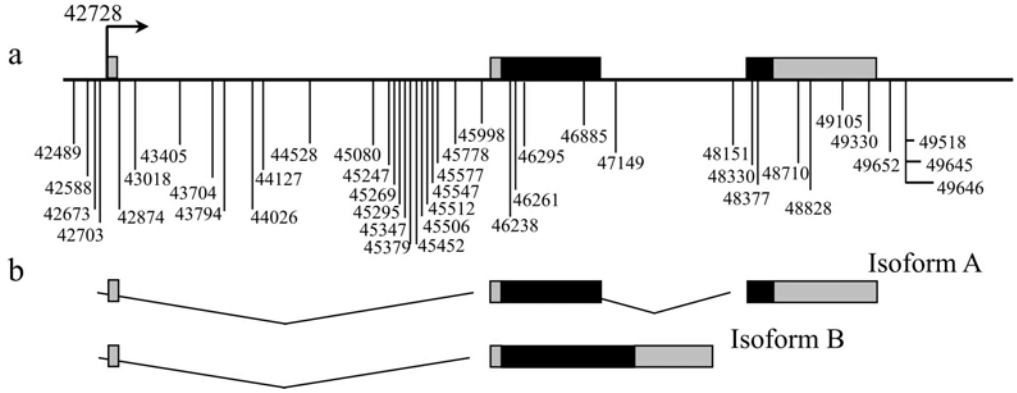


Figure 1. A schematic representation of the CCR2 gene with known mRNAs and polymorphisms. *a* – the region of the gene with polymorphisms whose positions correspond to the contig nucleotide sequence with accession number NT_079509. Arrow points to the transcription start, coding parts of exons are black, the 5′UTR and 3′UTR are shaded, *b* – the known tissue-specific mRNAs, isoforms A and B. The coding parts of exons are black, the 5′UTR and 3′UTR are shaded, and the introns removed by splicing are shown by thin line.

The nucleotide sequences flanking the different SNPs in the CCR2 gene, which we collected in the database, will make it possible to perform an *in silico* search of the polymorphism causing predisposition to MI. The functional analysis of the polymorphisms in the noncoding parts of the gene, using special software tools is required. The technology would predict the potential transcription binding sites, splicing sites and RNA secondary structures. The functional analysis of the SNPs that causes nonsynonymous substitutions in the protein can be combined with the software tools usually used to predict protein 3D structures and identify the functional motifs in protein. Two CCR2 mRNAs were detected. The isoform A includes the three exons of the CCR2 gene. The isoform B contains the first, second exons, and part of the second intron. Interestingly, the second intron differs from the other part of the CCR2 gene by low density of the polymorphisms. The polymorphisms are located on the flanks of this intron, not on its central part. It is possible that certain SNPs in the sequences of the flanks are involved in splicing regulation. Using contextual DNA analysis, we expect that computer-assisted data would clarify whether the polymorphisms are involved in regulation of the CCR2 gene expression. As a result, each polymorphism will be assigned a weight score and each will be ranged according to the priority of its putative effect on the final phenotype trait (MI).

The high cost of genotyping raises the question, how to choose the best – the less voluminous and most informative set of SNP polymorphisms for an associative survey of candidate genes or loci on a chromosome. It would appear that preliminary SNPs

weighing according to their potential functional contributions to a particular trait (disease) would allow us to elaborate a new algorithm for search of causative polymorphisms which predispose to common complex diseases. Thus, the proposed database for the human CCR2 gene polymorphisms contains informative guidelines for *in silico* search of the polymorphisms relating to diseases associated with the immune system responses, in particular those causing MI predisposition.

ACKNOWLEDGEMENTS

Work was supported in part by International Science and Technology Center (grant No. 2311), by the program "Dynamics of the gene resources of plant, animals and human" of the Russian Academy of Science and innovation project of Federal Agency of Science and Innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)". The authors are grateful to Lokhova I.V. for assistance in retrieval of the full text of publications.

REFERENCES

Bjarnadottir K. *et al.* (2005) Examination of genetic effects of polymorphisms in the MCP-1 and CCR2 genes on MI in the Icelandic population. *Atherosclerosis* (Epub ahead of print).

Gonzalez P. *et al.* (2001) Genetic variation at the chemokine receptors CCR5/CCR2 in myocardial infarction. *Genes Immun*., **2**(4), 191–195.

Le Y. *et al.* (2004) Chemokines and chemokine receptors: their manifold roles in homeostasis and disease. *Cell. Mol. Immunol*., **1**(2), 95–104.

Ortlepp J.R. *et al.* (2003) Chemokine receptor (CCR2) genotype is associated with myocardial infarction and heart failure in patients under 65 years of age. *J. Mol. Med*., **81**(6), 363–367.

Petrkova J. *et al.* (2003) CC chemokine receptor (CCR)2 polymorphism in Czech patients with myocardial infarction. *Immunol. Lett*., **88**(1), 53–55.

Voevoda M.I. *et al.* (2002) Association of the CCR2 chemokine receptor gene polymorphism with myocardial infarction. *Dokl. Biol. Sci*., **385**, 367–370.

LENGTH OF EXONS AND INTRONS IN GENES OF SOME HUMAN CHROMOSOMES

Atambaeva S.A., Ivashchenko A.T.*, Khailenko V., Boldina G., Turmagambetova A.
al-Farabi Kazakh National University, Almaty, 050038, Kazakhstan
* Corresponding author: e-mail: a_ivashchenko@mail.ru

Key words: exon, intron, gene, genome, *Homo sapiens*

SUMMARY

Motivation. Length and number of introns in genes of different eukaryotes, including human, varied within wide range of limits. It was important to clarify a quantitative regularity is in exon-intron organization of genes. The elucidation of exon and intron lengths variation in genes will promote determining intron function.

Results. The number of introns in genes was proportional to total length of exons and gene length of chromosomes 1, 2, 13, 19, 21 and 22 in *Homo sapiens* genome. The variations of intron and exon lengths in genes depended on number of introns in genes and genes density of DNA region.

INTRODUCTION

Genes containing introns were more than 90 % in nuclear genomes of *H. sapiens* (Venter *et al*., 2001). There was a considerable heterogenity of exon and intron lengths in genes, which provided determination of regularities of exon and intron lengths variability in every chromosome of *H. sapiens* genome. The number of genes including introns, number of introns in genes and a ratio of exon and intron length varied for different organisms (Duetsch, Long, 1999; Ivashchenko, Atambaeva, 2004). The relationship between exon and intron lengths that depend on number of introns in genes and gene density of DNA region in some chromosomes of *H. sapiens* has been determined.

METHODS

Nucleotide sequence of DNA have been extracted from GenBank (http://www.ncbi.nlm.nih.gov/). DNA sequentially was divided into regions of 1 Mbp length, which were put according to gene amount in groups from 1–11, 12–20, 21 and more genes per 1 Mbp. In each group have been analyzed the samples of genes containing 1–2, 3–5, 6–9, 10–14, 15 and more introns. The average values of intron and exon lengths, and total length of gene have been determined for each sample of genes. The analysis of frequency of occurrence of exon lengths has been made for following length intervals: 1–20, 21–40, 41–60 nt and so up to 400 nt and also more than 400 nt.

RESULTS

The allocation of genes along a DNA of chromosome 1 was heterogeneously also gene amount of region 1 Mbp length varied from zero point to 68 genes. In the group including 1 to 11 genes per region of chromosome 1 (average value was 4 genes/Mbp) exon length decreased from 282 to 135 nt, as well as the number of introns in genes (N_{in}) increased. The average total exon lengths (L_{ex}) in genes increased from 691 to 3163 nt and the positive correlation between N_{in} and L_{ex} variations was found out. This relationship was described by the following equation: $N_{in} = aL_{ex} + b$, where *a* and *b* are coefficients of linear regression. The values *a* and *b*, and coefficient of correlation (r) were shown in the Table1. The average gene length (L_{gn}) containing 1–2 introns was 22485 nt and it was 146296 nt from sample of genes containing 15 and more introns. There was a positive correlation between gene length and number of introns, which was represented by an equation: $N_{in} = cL_{gn} + d$, where *c* and *d* are coefficients of linear regression (Table 1).

Table 1. Parameters of linear regressions between number of introns and length of genes or sum of exon lengths

Genes/	Parameters of linear regressions						
1 Mbp	*a*	*b*	r	*c*	*d*	r	Nu.genes
			Chromosome 1				
4	0.0085	-4.06	0.997	0.00018	-3.96	0.966	273
16	0.0083	-3.40	0.997	0.00025	-0.70	0.967	325
26	0.0079	-3.88	0.989	0.00043	-1.64	0.991	320
32	0.0079	-3.74	0.997	0.00066	-2.12	0.971	396
			Chromosome 2				
4	0.0078	- 3.15	1.000	0.00016	- 2.93	0.984	428
4	0.0072	- 3.26	0.998	0.00013	- 0.48	0.991	525
15	0,0058	- 0,71	0.983	0,00024	- 0,39	0.985	376
29	0.0076	- 2.70	0.998	0.00060	- 2.16	0.964	186
			Chromosome 13				
3	0.0082	-7.11	0.987	0.00008	0.91	0.983	222
15	0.0088	-4.92	0.988	0.00023	-0.24	0.970	72
			Chromosome 19				
5	0.0093	- 4.39	0.994	0.00043	-4.10	0.861	34
16	0.0088	- 9.89	0.886	0.00030	-2.36	0.828	83
31	0.0080	- 4.99	0.988	0.00057	- 2.43	0.998	647
35	0.0068	- 2.65	0.988	0.00053	- 0.65	0.998	644
			Chromosome 21				
4	0.0070	- 1.90	0.997	0.00022	-3.94	0.986	110
17	0.0088	- 5.33	0.977	0.00042	- 2.52	0.961	100
30	0.0069	- 1.92	0.956	0.00053	- 7.63	0.952	18
			Chromosome 22				
5	0.0061	- 0.06	0.995	0.00013	1.33	0.972	91
15	0.0069	- 1.71	0.976	0.00034	- 2.83	0.992	124
28	0.0085	- 3.42	0.998	0.00047	- 2.49	0.987	273

It was established the change of the average exon length, when the number of introns in genes increased. For example, the average exon length decreased from 274 to 135 nt in 16 genes/Mbp group, sum of exon lengths increased from 706 to 2946 nt, length of genes increased from 5108 to 77198 nt accordingly for 1–2 introns genes and for genes containing 15 and more introns. The positive correlation between the sum of exon lengths and the number of introns in genes is shown (Table 1). The average intron length of the first gene group was 10576 nt and for the second gene group was 4128 nt. The result of the decrease of intron length was the contraction of the average gene length for all gene samples and accordingly a variation of linear regression parameters between gene length

and intron amount in genes (Table 1). While further increasing the gene density per 1 Mbp this tendency was observed too (Table 1). For example, in a gene group, where the density was 32 genes/Mbp, the average exon length decreased from 304 to 144 nt, the sum of exon lengths increased from 745 to 3308 nt, the gene length increased from 3918 to 32856 nt accordingly in 1–2 intron genes and in genes containing 15 and more introns. The relationship between the number of intron in genes and the total exon length for genes of four groups from chromosome 1 were shown in a Fig. 1. The correlation coefficients have been obtained from the great samples of genes and testify to a high reliability of this relationship ($p < 0.001$).

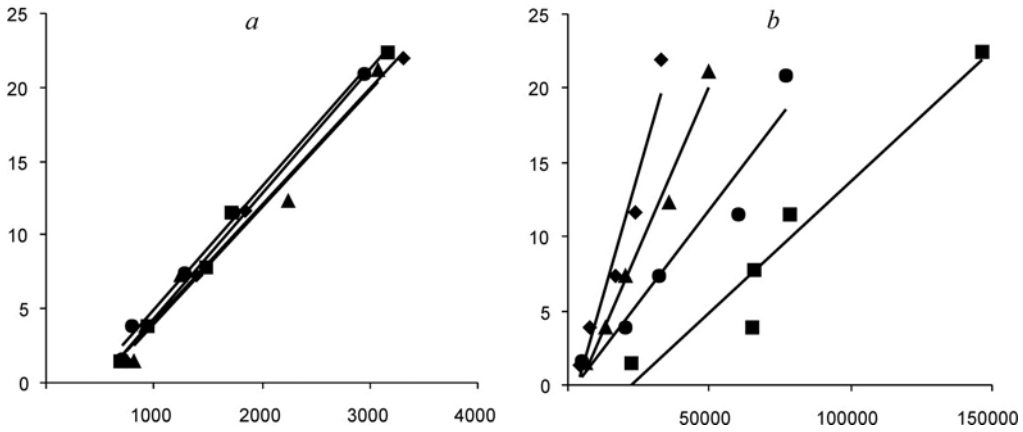


Figure 1. Correlations between total exon length (*a*), gene lengths (*b*) and number of introns in genes of chromosome 1. Regions having of gene density: 4 genes/Mbp – ■, 16 genes/Mbp – ●, 26 genes/Mbp – ▲ and 32 genes/Mbp – ♦; x-axis – sum of exon lengths (*a*) and gene lengths (*b*), nt; y-axis – number of introns in genes.

The greatest average density of genes/Mbp has chromosome 19 and two gene groups were formed a high gene density (Table 1). In both gene groups the relationship between sum of exon lengths and number of intron in genes was similar and was characterized by high correlation coefficients. Chromosome 13 has the lowest average density of genes/Mbp, however in two groups of genes the relationship between sum of exon lengths and number of introns in genes was similar and the high correlation coefficients were also presented too (Table 1). In the group with low gene density (3 genes/Mbp) the gene lengths increased from 27194 nt (1–2 introns in a gene) to 332554 nt (15 and more introns in a gene). The chromosomes 2, 21 and 22 had essential heterogeneity of gene distribution along a DNA. In all groups of genes between the sum of exon lengths and the number of introns in genes the relationship clearly appeared and had a high correlation coefficient (Table 1). The value of parameter *a* was similar for linear regressions of all the gene groups of every chromosomes. It obvious, the revealed connection is universal for all investigated human chromosomes and reflects an unknown intron function as sharing the protein coding part of a gene into segments.

The exon and intron share in the range of length 1–400 nt and more than 400 nt changed depending on gene sample in all the gene groups. In genes of *H. sapiens* chromosomes 1, 2, 13, 19, 21 and 22 the share of exons having length more than 400 nt decreased when increasing of number of introns in a gene, thus the share of exon having length 60–180 nt increased. For example, in the chromosomes 1 and 13 the share of exon with the length of more than 400 nt in 1-2 introns genes was 27.2 and 31.0 %, and in genes containing 15 and more introns 2.1 and 2.8 % accordingly (Fig. 2). The obtained data testify to the fact, that the genes having different intron number and located in different gene density regions have no the same exon-intron organization. The tendency

of increasing the number of intron in a gene, and the sum of exon lengths increased, testify to correcting function of introns on while unknown gene properties.

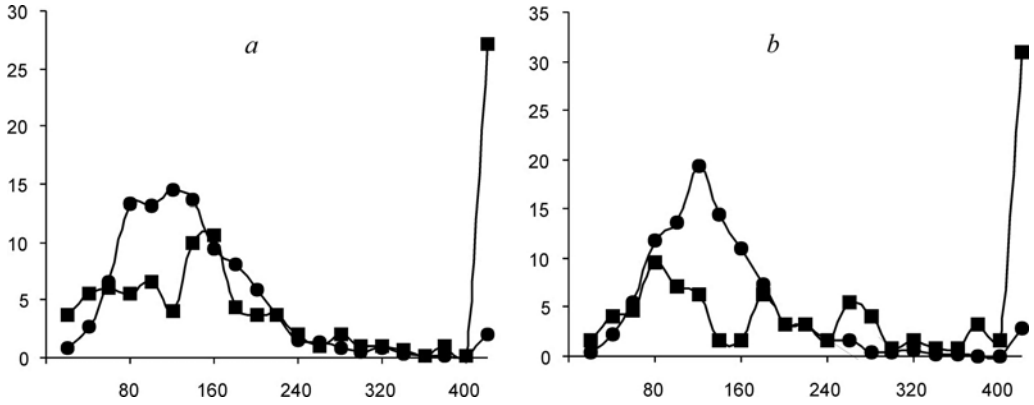


Figure 2. Variation of exon lengths in genes of the chromosome 1 (*a*) and chromosome 13 (*b*): ■ – exons lengths in 1–2 introns genes; ● – exons lengths in genes with 15 and more introns. x-axis – exon lengths, nt; y-axis – share exons, %.

REFERENCES

Venter J.C., Adams M.D., Myers E.W. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Duetsch M., Long M. (1999) Intron-exon structure of eukaryotic model organisms. *Nucl. Acids Res.*, **27**, 3219–3228.

Ivashchenko A., Atambaeva S. (2004) Variation in lengths of introns and exons in genes of the *Arabidopsis thaliana* nuclear genome. *Rus. J. Genet.*, **40**, 1179–1181.

THE EXON AND INTRON LENGTHS IN *ARABIDOPSIS THALIANA* AND *CAENORHABDITIS ELEGANS* GENES

Atambaeva S.A., Ivashchenko A.T.*
al-Farabi Kazakh National University, Almaty, 050038, Kazakhstan
* Corresponding author: e-mail: a_ivashchenko@mail.ru

Key words: exon, intron, gene, genome, *A. thaliana*, *C. elegans*

SUMMARY

Motivation. The variability of intron lengths and intron numbers in genes of various organisms is very different and detection of exon-intron regularity requires investigation of the gene organization. It is necessary to elucidate exon-intron structures of genes of various genomes for clearing up biological role of introns.

Results. The number of introns in a gene is proportional to the sum of exon lengths and length of genes in *A. thaliana* and *C. elegans* genomes. The changes of exon and intron lengths in *A. thaliana* and *C. elegans* genes possess features depending on intron number in genes and a gene density of DNA.

INTRODUCTION

Genomes of *C. elegans*, *A. thaliana*, *D. melanogaster*, *O. sativa, H. sapiens* and other eukaryotes have more than 85 % genes with introns. There is the considerable heterogeneity of exon and intron lengths, which promotes detection of variability of exon and intron lengths regularities in the genomes of these organisms (Duetsch, Long, 1999). In different organisms the intron number in genes, a ratio between exon and intron lengths, the number of genes with introns, etc. variate. The purpose of the present work is investigation of relationship between intron number in genes, exon and intron lengths, a gene density of DNA all chromosomes of *A. thaliana* and *C. elegans*.

METHODS

Nucleotide sequences of DNA of *A. thaliana* and *C. elegans* genomes have been extracted from GenBank (http://www.ncbi.nlm.nih.gov/). Genes of each chromosome containing 1, 2, 3, 4, 5, 6–10, 11–15, 16 and more introns were divided into groups. The intron and exon lengths, the sum of exon lengths and gene lengths in these groups have been determined. A frequency of occurrence of intron and exon lengths have been analyzed in the ranges 1–20, 21–40, 41–60 nt up to 400 nt as well as more than 400 nt.

RESULTS

The share of one intron genes in all chromosomes of *A. thaliana* was about 20 %. If intron number in genes increased, the share of such genes decreased. The genes with 11–15 intron genes were less than 1 %. One intron genes of *C. elegans* were about 10 % and the share of genes containing more intron number (2, 3, 4 and 5) gradually increased to14 % and then sharply decreased for genes containing 6 and more introns. The sum of exon lengths was 76–80 % (*A. thaliana*) and 63–74 % (*C. elegans*) in one intron genes and that was 46–50 % and 36–40 % in 11–15 intron genes accordingly. The correlation between changes exon lengths, intron lengths and intron number in *A. thaliana* genes have been established. The exon and intron lengths decreased if intron number increased. The intron number was proportional to the sum of exon lengths or gene lengths with high correlation coefficient (r) in genes of *A. thaliana* and *C. elegans* (Fig. 1).

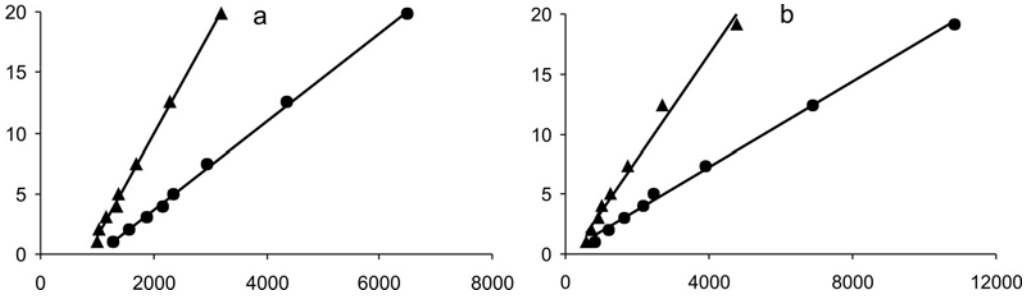


Figure 1. The correlation between exon lengths sum (▲), gene lengths (●) and intron numbers in genes of *A. thaliana* chromosome 4 (*a*) and *C. elegans* chromosome II (*b*). x-axis – gene lengths and exon lengths sum, nt; y-axis – intron numbers in genes.

These dependencies were described by the following equations: $N_{in} = aL_{ex} + b$ and $N_{in} = cL_{gn} + d$ accordingly where N_{in} is intron number in a gene, L_{ex} is exon lengths sum, L_{gn} is gene length, *a*, *b*, *c* and *d* are parameters of a linear regressions. Magnitudes of these parameters are given on Table 1. The received data testify to the relationship between exon lengths sum, genes lengths and changes of intron numbers in both organisms at increasing of intron numbers in genes. The high correlation coefficients between change of exon and intron lengths have been observed for *A. thaliana*: $L_{in} = mL_{ex} + n$ (Table 1). This regularity was similar for genes of all *A. thaliana* chromosomes. Such kind of correlation was missing in *C. elegans* genes (Table 1).

The gene number of a double-stranded DNA region of 0.3Mbp length varied from 8 to 99 for *A. thaliana* and from 26 to 110 for *C. elegans*. The relationship was similar between changes of *A. thaliana* gene lengths, exon lengths sum and intron numbers at the density of 86 and 27 genes/0.3Mbp consequently:

$N_{in} = 0.0119L_{ex}$-9.59; r = 0.939; $p < 0.001$ и $N_{in} = 0.0051L_{gn} – 5.06$; r = 0.993; $p < 0.001$;

$N_{in} = 0.0066L_{ex}$-3.00; r = 0.987; $p < 0.001$ и $N_{in} = 0.0034L_{gn} – 2.97$; r = 0.991; $p < 0.001$.

Intron lengths were 4 times less in genes from the high gene density regions (99 genes/0.3Mbp), than intron lengths of genes at low gene density regions (38 genes/0.3Mbp) in *C. elegans* genome. The relationships between changes of gene lengths, exon lengths sum and intron numbers were different at the density of 86 and 27 genes/0.3Mbp:

$N_{in} = 0.0057L_{ex} – 2.6$, r =0.994, $p < 0.001$ and $N_{in} = 0.0029L_{gn} – 1.12$, r = 0.997, $p < 0.0005$;

$N_{in} = 0.0039L_{ex} – 0.7$, r = 0.997, $p < 0.0005$ and $N_{in} = 0.0008L_{gn} – 1.04$, r = 0.992, $p < 0.001$.

The average gene lengths and ratio of exon and intron lengths were the same in *C. elegans* regions containing the close gene numbers (63÷65 genes/0.3Mbp) with the

different GC-content (39 % and 34 %). Neither GC-content nor gene density didn't influence on the ratio of exon and intron lengths in *A.thaliana* genome, although regions were different 7 times by a gene density.

Table 1. The parameters of correlation of gene lengths, exon lengths sum and intron number; exon lengths and intron lengths of *C. elegans* and *A. thaliana* genes

Chromo-some	Parameters				Parameters			
	a	*b*	r	p	*c*	*d*	r	p
C. elegans								
I	0.0045	-0.94	0.998	<0.0001	0.0018	-0.41	0.999	<0.0001
II	0.0049	-1.29	0.997	<0.0005	0.0018	-0.01	0.998	<0.0001
III	0.0040	-0.28	0.991	<0.001	0.0015	-0.37	0.997	<0.0005
IV	0.0046	-0.84	0.995	<0.001	0.0017	-0.04	0.999	<0.0001
VI	0.0051	-1.54	0.996	<0.001	0.0018	-0.18	0.995	<0.001
X	0.0057	-1.27	0.997	<0.0005	0.0023	-0.71	0.999	<0.0001
A. thaliana								
1	0.0087	-6.83	0.994	<0.001	0.0036	-3.21	0.997	<0.0005
2	0.0084	-6.22	0.987	<0.001	0.0038	-3.41	0.995	<0.001
3	0.0088	-6.55	0.995	<0.001	0.0037	-3.08	0.998	<0.0001
4	0.0084	-6.93	0.999	<0.0001	0.0037	-3.70	0.999	<0.0001
5	0.0099	-8.03	0.998	<0.0001	0.0042	-3.99	0.999	<0.0001

A. thaliana					*C. elegans*				
Chr.	m	n	r	p	Chr.	m	n	r	p
1	0.34	106	0.987	<0.001	I	0.25	287	0.172	>0.1
2	0.43	87	0.996	<0.001	II	-0.41	370	-0.312	>0.1
3	0.25	120	0.960	<0.01	III	0.10	316	0.061	>0.1
4	0.42	100	0.956	<0.01	IV	-0.57	429	-0.352	>0.1
5	0.37	87	0.995	<0.001	V	-0.64	402	-0.528	>0.1
					X	-0.66	386	-0.618	>0.05

The exon and intron length varied specifically in the intervals of 1–400 nt length and more than 400 nt for all *A. thaliana* gene groups. The share of exons with more than 400 nt length was 5 % in chromosome 1 genes containing 1–15 introns, i.e. 9 times less, than in one intron genes (44 %). The share of exons with the length from 60 to 180 nt increased from 23 % to 76 % (Fig. 2*a*).

The shares of introns with the length from 80 to 120 nt were 40 % in one intron genes and 68 % in 11–15 intron genes. The shares of introns with length more than 400 nt were 20 % and 5 % accordingly (Fig. 2*b*). While exon number increasing in genes, a redistribution of the intron lengths was the result of reduction of intron share with length more than 400 nt and increasing of its number for an interval from 80 to 120 nt without change of the intron share in the intermediate intervals of lengths. The intron share for a length from 140 to 400 nt was similar: 26÷35 % in genes of all *A. thaliana* chromosomes.

The share of exons with length more than 400 nt decreased, when the intron number in a gene increased and the share of exons with length from 60 to 180 nt increased simultaneously too in genes of all *C. elegans* chromosomes. The share of exons with length more than 400 nt was 19 % in one intron genes and 9 % in 11–15 intron genes. Thus the share of exons with length from 60 to 180 nt increased from 46 to 60 % in chromosomes IV (Fig. 2*c*). The intron lengths didn't depend on intron numbers in *C. elegans* genes (Fig. 2*d*).

The recieved data show that the exon-intron organization wasn't the same for genes containing various intron numbers and located in the different chromosome regions. The clearly expressed regularity testify to correcting role of introns for unknown gene features, because intron number increased when the exon lengths sum increased.

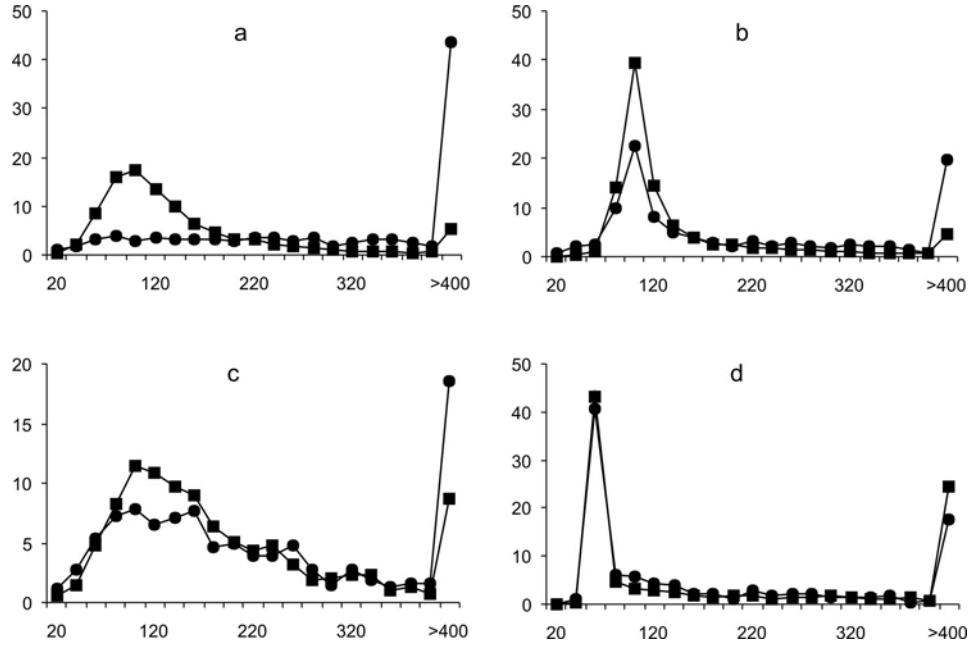


Figure 2. Variation of exon and intron lengths in genes of *A. thaliana* chromosome 1 (*a*, *b*) and of *C. elegans* chromosomes IV (*c*, *d*): *a* – exons of one intron genes (-●-) and of 11–15 introns genes (-■-); *b* – introns of one intron genes (-●-) and of 11–15 introns genes (-■-); *c* – exons of one intron genes (-●-) and of 11–15 intron genes (-■-); *d* – introns of one intron genes (-●-) and of 11–15 intron genes (-■-). x-axis – exon and intron lengths, nt; y-axis – share of exons and introns, %.

REFERENCES

Duetsch M., Long M. (1999) Intron-exon structure of eukaryotic model organisms. *Nucl. Acids Res.*, **27**, 3219–3228.

STATISTICAL CHARACTERIZATION OF CONSERVED NON-CODING ELEMENTS IN VERTEBRATES

***te Boekhorst R.*[*1], *Walter K.*[2], *Elgar G.*[3], *Gilks W.R.*[2], *Abnizova I.*[2]**
[1] University of Hertfordshire, College Lane, Hatfield, UK; [2] MRC-BSU, Robinson Way, Cambridge, UK;
[3] Queen's Mary College, London, UK
[*] Corresponding author: e-mail: r.teboekhorst@herts.ac.uk

Key words: CNE-elements, non-coding DNA, regulatory DNA

SUMMARY

Motivation: Recently, a set of highly Conserved Non-coding Elements (CNE's) was derived from a *Fugu*-human genome comparison. We characterise some statistical features common to these elements in order to facilitate their identification *in silico*.

Results: We found a pronounced pattern around the borders of CNEs: GC-rich flanking regions of low entropy compared to AT-rich, high entropy sequences within the borders. We also identified the most abundant significant motifs inside and adjacent to the borders of CNE's. At the borders, motifs are significantly clustered which points to their possible role as binding sites.

INTRODUCTION

Only around 1.2 % of human DNA is known to be coding for proteins. Our knowledge of the role and location of other elements is limited and new types of sequences of unknown function are still discovered. Recently, several sets of highly conserved non-coding sequences have been identified in vertebrate genomes (Woolfe *et al*., 2004; Bofelli *et al*., 2005; Dermitzakis *et al*., 2005). A combination of comparative genomic studies and laboratory experiments has shown that these conserved non-coding elements (CNEs), most of which are more conserved than protein-coding exons, may be regulatory elements (Moses *et al*., 2005; Xie *et al*., 2005).

Conserved regulatory regions have been the objects for motif discovery by phylogenetic foot-printing algorithms (Blanchette, Nompa, 2002). However, most efforts been related to the promoter motifs (FitzGerald *et al*., 2004) and although CNEs appear to have striking "signatures" (Walter *et al*., 2005), little motif discovery has been done for CNEs.

Here, we focus on the motif identification and statistical characterization of the CNEs collected by Woolfe *et al*. (2004). Based on a MEGABLAST comparison between human and pufferfish (*Fugu ribripes*) genomes, they identified about 1400 highly conserved non-coding sequences. Most of these sequences are located in and around developmental regulation genes and when some of them were tested in the laboratory, they appeared to drive tissue-specific gene expression in early development (Woolfe *et al*., 2004). These facts encouraged us to consider CNEs as putative regulatory regions, namely enhancers, and to check whether they could be characterised by some of the basic statistical properties of regulatory regions such as the abundance (Papatsenko *et al*., 2002) and the typical spatial distribution (FitzGerald *et al*., 2004) of binding motifs.

MATERIALS AND METHODS

The set of CNEs identified by Woolfe et al. (2004) contains 1373 elements, vary in size from 53 bp to 740 bp (mean length 200 bp) and a level of conservation is from 68 % to
98 % identity. We use this data set to build up a CNE lexicon, and check for presence of statistical properties typical of cis-regulatory regions. To characterise CNE borders, we generated two positive data sets of 5′ and 3′ CNE flanking regions of 50 bp each (upCNE and CNEdown) from the 1231 CNEs which are longer than 100 bp.

Likelihood of motifs. It is known that certain sequences that operate as "binding motifs" are surprisingly abundant within regulatory regions. Given the DNA composition of a region, the globally most abundant motifs may be defined as those that are most likely to occur. In the work reported here, we determined the motifs in the CNE-flanking alignments with the highest likelihood. We will show that these motifs are more abundant than expected due to the composition of the upCNE and downCNE regions.To do so, we generated a large number (10000) of "surrogate" alignments with the same position-dependent composition as the CNE alignment under consideration by randomly shuffling the original sequence 10000 times. Next, Z-scores of words from the original lexicon were calculated as standardized deviations from the mean frequency (of the same words) of the randomised alignments. Words from the original CNE alignment are defined as significant if their Z-score exceeds 2 standard deviations

Spatial Distribution (clustering) of Words. Some words, not necessarily the most frequent in the CNE alignments, could be functionally important as binding sites and could therefore be clustered around CNE borders. We use the local frequency of words (i.e. within columns) in the alignment to determine their degree of clustering. To assess the statistical significance of word clustering, the clustering coefficient, CC, is defined for each word x_i in each start position j and sequence (we omit the indexing of sequence for simplicity) as $CC^j(x_i) = \frac{N^j(x_i) - \overline{N(x_i)}}{\sigma(x_i)}$, where $N^j(x_i)$ is the occurrence of word x_i starting in position j, $\overline{N(x_i)}$ is the mean frequency (i.e. of $N^j(x_i)$) over all positions j in the alignment, and $\sigma(x_i)$ is the standard deviation of $N^j(x_i)$. A word x_i is significantly locally clustered (or anti-clustered) in position j, if $|CC^j(x_i)| > 2$.

Compositional homogeneity. While aligning CNEs, we had the impression that the flanking regions vary stronger in composition than the CNEs themselves. We estimated the di-nucleotide entropy separately in each 50 bp flanking region and each 50 bp CNE sequence of the alignments as a quantitative measure for their compositional diversity. In previous work, we have shown that the entropy of regulatory regions is intermediate between that of coding (highest entropy) and non-coding, non-regulatory regions (lowest entropy) (Orlov *et al*., 2006).

Sequential persistence. We used the Hurst exponent as a measure of the stationarity of the DNA sequence around the CNEs border. The Hurst coefficient was calculated by Rescaled Range Analysis. We applied this method by transforming a DNA sequence into a binary code of $x_k = +1$ for $k = G, C$ and $x_k = -1$ for $k = A, T$ (Orlov *et al*., 2006). In case of random, identical and independent occurrences of nucleotides in DNA, H equals 0.5. A high Hurst exponent (> 0.5) points to extensive autocorrelations (i.e. non-stationarity). A series that contains a significant change in composition is therefore expected to be characterized by H > 0.5. In previous work, we have shown that the Hurst exponent of regulatory regions (H ~ 0.62) is intermediate between that of coding (H < 0.5, indicating anti-persistence) and non-coding, non-regulatory regions (H ~ 0.67) (Orlov *et al*., 2006).

To statistically characterise the CNEs and their borders, we calculated Entropy and Hurst exponent between the following regions. For entropy: 50bp upstream flanking

regions, the first- and last 50 base pairs of the CNE itself and 50 bp downstream flanking regions. The entropy values of these stretches were compared to sequences (50 bp long) of randomly picked non-coding, non-regulatory DNA in *Fugu*. For the Hurst exponent: an upper CNE bordering region containing the first 50 base pairs before and after the CNE start position and a lower CNE bordering region consisting of the last 50 base pairs of a CNE and the first 50 base pairs after the stop-position of that CNE. We compared Hurst exponent values of the two bordering regions with those of randomly selected non-coding, non-regulatory DNA within the same window length as the aligned CNEs.

RESULTS

We aligned the highest scoring significant words (of length 12 here) with respect to their start positions in columns, and put them into to WebLogo format (Crooks *et al*., 2004). The "logos" of these over-represented words are shown in Fig. 1. The lowest scoring words are visualized in Fig. 2. As one can see from Fig. 1–2, over- and under-represented words are AT rich and CG poor respectively and their start positions appear to cluster close to the CNE borders.

Figure 1. Left: highest scoring 12-mers in CNEdown, the CNE border is at position 50. Right: high scoring 12-mers in upCNE, the CNE border is at position 51.

Figure 2. Left: lowest scoring 12-mers in CNEdown, the CNE border is at position 50. Right: lowest scored 6- and 12-mers in upCNE, the CNE border is at position 51.

We found that there this clustering (CC > 2) of certain short patterns around the CNE borders is significant. This is shown in the plot of start-positions of significantly large clustering coefficients (> 2) along the alignment (Fig. 3).

The average entropy of upstream and downstream flanking regions (both E = 2.43) is significantly lower than that of regions within the CNE border (both E = 2.52) (Newmans-Keuls ad hoc comparisons after a 1 way repeated measurement ANOVA within sequences. A Friedman test – as a non-parametric alternative to the repeated measurement ANOVA – backed up the results). The entropy of non-coding, non-regulatory regions (E = 2.49) is significantly higher than that averaged over the two flanking regions (t-Test: t = 8.52, df = 2460, $p < 0.0001$) and significantly lower than that averaged over the two within CNE regions (t-Test: t = 5.82, df = 2460, $p < 0.0001$). Mann-Whitney U tests, as non-parametric alternatives to the parametrical t-Tests demonstrated a significant difference between the non-coding, non-regulatory DNA on the one hand and the combined (= averaged) data of the two flanking regions on the other hand, but not with the combined (averaged) within CNE sequences.

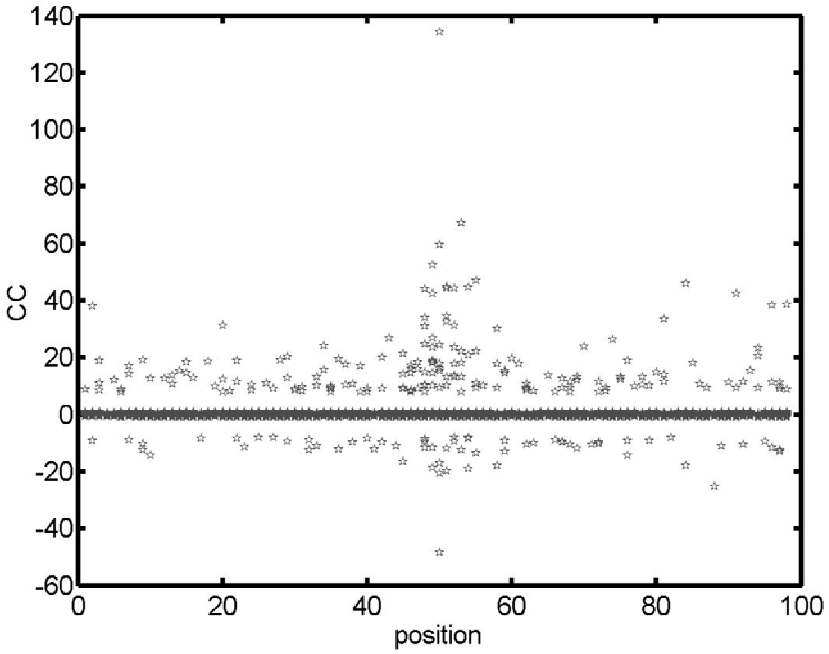


Figure 3. The clustering of four mers around the upCNE border, at position 50 bp. The vertical axis shows modified CC (= sign(CC)*CC2) for visualization purposes. The horizontal axis shows position in the upCNE alignment.

The Hurst exponents between upper and lower bordering regions (respectively H = 0.65 and 0.66) do not differ significantly, but do so between the values of each of the two bordering regions and those of non-coding, non-regulatory regions (H = 0.59) (Newmans-Keuls ad hoc comparisons after a 1 way ANOVA on log-transformed data). Together, these results point to a change in composition at the borders of CNEs.

DISCUSSION AND CONCLUSION

We have showed that the motifs around CNE borders are not just the consequence of compositional bias. In addition, we identified the following statistical "signatures" of CNEs: (i) the sequences around CNE borders are surprising rich in globally and locally over-represented motifs; (ii) CNE borders appear to correspond to a pronounced change-point in composition; (iii) flanking CNE sequences have low entropy and are CG rich whereas the CNE themselves are AT-rich, and have a higher entropy compared to the flanking regions. Although it has been put forward that some CNEs might be matrix attachment regions (Glazko *et al.*, 2003) or participate in inter-chromosomal interactions (Muller, Schaffner, 1990), due to their statistical properties CNEs might indeed function as regulatory regions. They contain more statistically significant abundant words than expected by chance, many words are clustered close to their start positions and their entropy is on average 2.52, i.e. in between that of typical coding regions (~2.68, refs) and non-coding, non-regulatory regions (2.48). These findings corroborate evidence in the literature (Hardison, 2000; Nobrega *et al.*, 2003). The most significantly clustered motifs could be candidates for TFBS cores. Note that some words are rarer than expected (anti-clustered) near the border (Fig. 3). They could be candidates for under-represented TFBS binding sites. Comparison with randomly picked non-coding non-regulatory (NCNR) DNA revealed that in the latter locally

highly clustered motifs are fairly uniformly spread over alignments, in contrast to the clustering around the borders of CNEs, which is typical for regulatory regions (FitzGerald *et al*., 2004).

REFERENCES

Blanchette M., Tompa M. (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res.*, **12**(5), 739–748.

Bofelli D., Nobrega M., Rubin E. (2005) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.*, **6**, 151–157.

Crooks G.E., Hon G., Chandonia J.M., Brenner S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Dermitzakis M., Reymond A., Antonarakis S. (2005) Conserved non-genic sequences-an unexpected feature of mammalian genomes. *Nat. Rev. Genet*., **6**, 151–157.

FitzGerald P., Shlyakhtenko A., Mir A., Vinson C. (2004) Clustering of DNA sequences in human promoters. *Genome Res.*, **14**, 1562–1574.

Glazko G., Koonin E., Rogozin I., Shabalina S. (2003) A significant fraction of conserved noncoding DNA in human and mouse consists of predicted matrix attachment regions. *Trends Genets*, **19**, 119–124.

Hardison R.C. (2000) Conserved non coding sequences are reliable guides to regulatory elements. *Trands Genets*, **16**, 369–372.

Muller H.P., Schaffner W. (1990) Transcriptional enhancers can act in *trans* binding. *Trends Genets*, **6**, 521–556.

Nobrega M.A., Ovcharenko I., Afzal V., Rubin I. (2003) Scanning human gene deserts for long-range enhancers. Science, **302**, 413.

Orlov Y.L., te Boekhorst R., Abnizova I. (2006) Statistical measures of the structure of genomic sequences: entropy, complexity, and position information. *J. of Bioinformatics and Computational Biology* (In press).

Papatsenko D., Makeev V., Lifanov A., Regnier M., Nazina A., Desplan C. (2002) Extraction of functional binding sites from unique regulatory regions: the *Drosophila* early developmental enhancers. *Genome Res.*, **12**(1), 470–481.

Walter K., Abnizova I., Elgar G., Gilks W.R. (2005) Striking Nucleotide Frequency Pattern at the Borders of Highly Conserved Vertebrate Non-Coding Sequences. *Trends in Genetics*, **21**, 438–440.

Woolfe A., Goodson M., Goode D., Snell P., Smith S., Vavouri T., McEwen G., Gilks W., Walter K., Abnizova I., Edwards Y., Elgar G. (2004) Highly conserved non coding sequences are associated with developmental control genes in vertebrates. *PloS Biology*, **3**, e7.

INTERPRETATION OF RESULTS OF SOM ANALYSIS OF MICROARRAY DATA BY PRINCIPAL COMPONENTS

***Efimov V.M.*[*1], *Badratinov M.S.*[2], *Katokhin A.V.*[2, 3]**
[1] Institute of Systematics and Ecology of Animals, SB RAS, Novosibirsk, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [3] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: vmefimov@ngs.ru

Key words: microarray data, SOM (self-organizing maps) analysis, PCA (principal components analysis), visualization

SUMMARY

Motivation: Microarray technology provides a massively parallel means to study gene expression on a global scale. There are many challenges associated with the analysis of microarray data due to its inherent complexity and high dimensionality. Although there is a diverse range of analytical techniques available for finding groups in gene expression data, clustering and partitioning are currently the key areas of microarray data mining. Combining the analytical techniques could provide new ways to improve grouping quality and interpretability.

Results: We applied the method of principal components to a united sample of gene expression profiles, presented by Borovecki *et al.* (2005), and the centers of SOM clusters that we calculated. This allowed us to give a meaningful interpretation to the clusters obtained.

INTRODUCTION

The Kohonen's self-organizing maps (SOM analysis) are among the methods widely used for analyzing microarray data (Kohonen, 1997). The essence of the method is a nonlinear transformation of a set of dots representing, for example, gene expression profiles from a space of a large dimension to a space of a small dimension with concurrent clustering of the dots (Tamayo *et al.,* 1999; Hand and Heard, 2005). However, regarding the centers of SOM clusters as new dots in the initial space brings about an interesting new possibility to analyze their layout among the initial set of dots.

For this purpose, we propose to apply the principal components analysis (PCA). This method as a variant of the projection methods used for analysis of microarray data is sometimes also called singular value decomposition (SVD; Alter *et al.*, 2000; Wall *et al.*, 2001).

The PCA applied to a united sample of gene expression profiles and centers of SOM clusters allows for a meaningful interpretation of the clusters obtained, thereby assisting the understanding of the potential and results of SOM analysis.

MATERIAL AND METHODS

A massive of microarray expression data obtained using GE Codelink Human Uniset I, II, and 20K (GPL1449) microarrays. Upon filtration procedure (the empty lines were removed), the massive contained 17 526 complete gene expression profiles (lines) from 31 samples of the peripheral blood (columns: 12 patients with Huntington's disease, 5 presymptomatic individuals, and 14 normal cases; Borovecki *et al*., 2005). The files with these experimental data as a set of GDS1332 data were extracted from the GEO database (Barrett *et al.*, 2005; http://www.ncbi.nlm.nih.gov/geo/).

The centering and normalization were performed by our own program.

The program Cluster 3.0 (de Hoon *et al.*, 2004) was used to perform the SOM and the PCA procedures.

The software package STATISTICA 6.0 was used for visualization.

RESULTS AND DISCUSSION

Upon a logarithmic transformation, the data massif was centered and normalized over the columns to eliminate nonuniformity within the samples and over the lines to remove the scale effects. Due to such transformation, all the dots are located on the surface of a 30-mer sphere (one degree of freedom disappears due to centering). According to the SOM algorithm, 30 centers of clusters were created and 1 000 000 iterations of approximation of these centers to the clusters of initial dots were performed followed by centering and normalization of these centers.

The set of initial dots supplemented with the set of SOM cluster centers were processed as a single sample by principal components analysis. The first three components together gave 26.23 % of the total variance. Fig. 1 shows the arrangement of samples on the plane of the first and third eigenvectors.

Fig. 2 demonstrates the arrangement of gene expression profiles and SOM cluster centers on the plane of the same principal components as well as locations of the profiles for 12 genes—markers of Huntington's disease selected based on the results of confirming real-time PCR experiments from 322 genes that displayed the most significant and pronounced differential expression between the groups of patients and healthy control in microarray experiments (Borovecki *et al.*, 2005).

All the 12 marker genes fell into two SOM clusters (2nd and 12th), forming a rather tight group. The centers of these clusters fell into virtually the same dot on the plane of the first and third principal components. As is evident from Fig. 1, these components are responsible for the differences between the patients with pronounced disease symptoms and healthy individuals. Note that the distinctions between clusters 2 and 12 manifest themselves in the second principal component (8.65 % of the total variance; data not shown), which is responsible for the deviation of presymptomatic patients from both the norm and Huntington's disease cases.

Presumably, the group of candidate marker genes for Huntington's disease may be expanded considerably with all the genes whose profiles are located near the centers of these clusters, thereby increasing the sensitivity and reliability of diagnostics of various Huntington's disease states.

Thus, the PCA applied to united sample of gene expression profiles and centers of SOM clusters allows for visualizing the location of SOM clusters and their centers among the rest set of objects as well as obtaining a meaningful interpretation of the clusters obtained. Both possibilities assist essentially the understanding of SOM analysis potential and results.

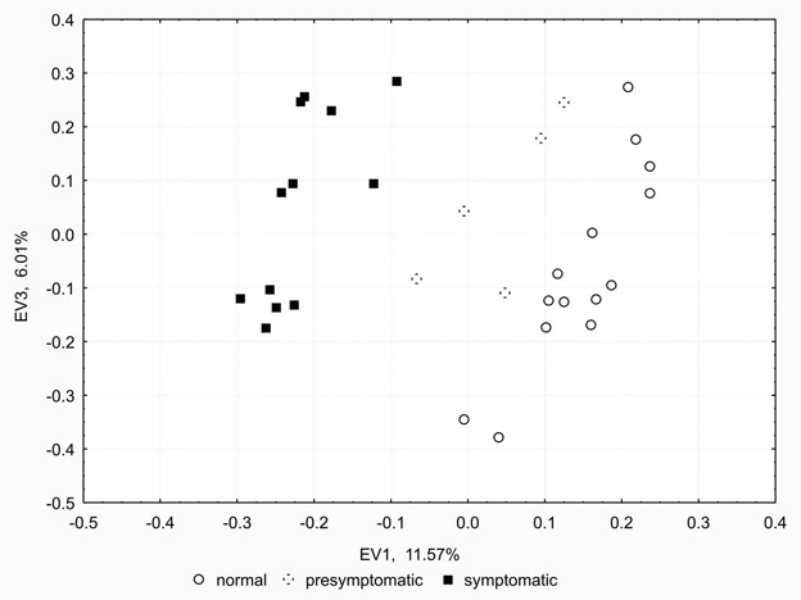


Figure 1. Arrangement of the objects on the plane of the first and third eigenvectors.

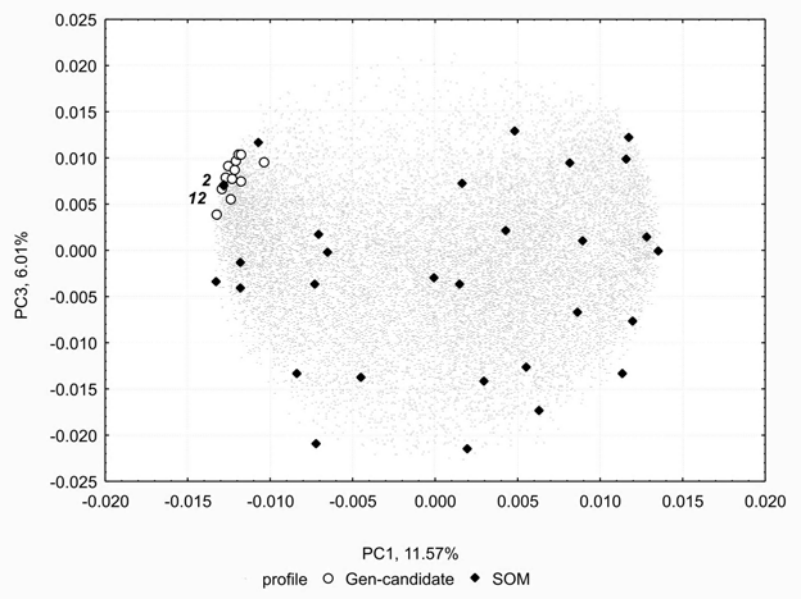


Figure 2. Arrangement of gene expression profiles and SOM cluster centers on the plane of the first and third principal components.

ACKNOWLEDGEMENTS

This work was supported by the innovation project of the Federal Agency for Science and Innovations IT-CP.5/001 “Development of software for computer modeling and design in postgenomic systems biology (systems biology *in silico*)”.

REFERENCES

Alter O. *et al*. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl. Acad. Sci. USA*, **97**(18), 10101–10106.

Barrett T. *et al.* (2005) NCBI GEO: mining millions of expression profiles--database and tools. *Nucl. Acids Res*., **33**(Database issue), D562–566.

Borovecki F. *et al.* (2005) Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc. Natl. Acad. Sci. USA*, **102**(31), 11023–11028.

de Hoon M.J. *et al.* (2004) Open source clustering software. *Bioinformatics*, **20**(9), 1453–1454.

Hand D.J., Heard N.A. (2005) Finding groups in gene expression data. *J. Biomed. Biotechnol*., **2005**(2), 215–225.

Kohonen T. (1997). *Self-Organizing Maps*. Springer, Berlin.

Tamayo P. *et al.* (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, **96**(6), 2907–2912.

Wall M.E. *et al.* (2001) SVDMAN-singular value decomposition analysis of microarray data. *Bioinformatics*, **17**(6), 566–568.

AN EXTENDED BACKUS-SYSTEM FOR THE REPRESENTATION AND ANALYSIS OF DNA SEQUENCES

Hofestädt R.
Bielefeld University, AG Bioinformatics, Bielefeld, Germany
e-mail: hofestae@techfak.uni-bielefeld.de

Key words: sequence analysis, backus-system, complexity

SUMMARY

Motivation: Using methods based on those of molecular biology isolating and sequencing of DNA and proteins is possible. This data must be stored using specific database systems which must be available via the internet. The reason to require an efficient implementation of these database systems is the exponential increase of their entries.

Results: In this paper we define an extended Backus-System which will allow an efficient representation of this data. Furthermore, important topics like the "complexity of life" can be discussed using this new formal representation.

INTRODUCTION

The structured storage of the sequenced biological data, the analysis of this data, and the availability of this data (deoxyribonucleid acid and proteins) requires the methods of computer science (Tanaka, 1992). The main tasks are to develop new database systems and efficient algorithms for the analysis of this data. Moreover, computer scientists have to support the usage of supercomputers. The electronic analysis of biological data is based on a suitable representation and implementation of this data. A well known method is the application of formal languages which was introduced by (Brendel, Busse, 1984). They used chomsky type-3 languages to describe biological data. The application of formal languages was extended by (Collado-Vides,1991), when he introduced the description of DNA functional units. In this paper we will discuss the grammatical formalization of nucleic acid. For this Brendel and Busse suggested regular expressions. However, this formalization is not able to satisfy all there requirements. Today it is known that any DNA sequence of any functional unit can be characterized by its specific lenght. Moreover, in any DNA functional unit sequence variations can appear which are based on the redundancy of the genetic code. Chomsky grammars are not able to express these specific features of each functional class. This is a reason for using an expanded version of the Backus-System.

GRAMMARS

A *chomsky type-0 grammar* is given by a quadrupel $G = (\Phi, \Sigma, P, A)$ (Maurer, 1977). Φ and Σ are finite alphabets with $\Sigma \cap \Phi = \varnothing$ and $\Phi \cup \Sigma = \Gamma$. The elements of Φ are called variable symbols and the elements of Σ are called terminal symbols. P denotes the set of all rules and the variable symbol A is called the axiom. P is a finite set of ordered pairs of

Γ^*. (Γ,P) is called *production system*. For each rule (u,v) $\in$ P u contains one or more variable symbols. For two words w,y $\in \Gamma^*$ we say that w could be derivated into y, in symbols w => y, iff . z_1, z_2 $\in\Gamma^*$ w = z_1uz_2 and y=z_1vz_2 and (u,v) ε P. w => y is called *one-step derivation* from w into y. Let w,y $\in \Gamma^*$, we call w derivable into y, in symbols w =>* y, if there exists a sequence of words w_0, ..., $w_n \in \Gamma^*$ (n>1) which represents the following one-step derivations w = w_0 => ... w_{n-1} => w_n. A sequence w_1, .., w_n is called *derivation* of the length n. L(G) = { x | x $\in \Sigma^*$ and A =>* x } is the set of all words which are derivable from the axiom A. L(G) is called *rule based language*. A rule is called *linear*, if it is of the form A $\rightarrow$ x By with A,B $\in \Phi$ and x, y $\in \Sigma$*. A rule is called *right-linear* (*left-linear*), if the rule is of the form A $\rightarrow$ xB (A $\rightarrow$ Bx). A rule is called closed, if it is of the form A $\rightarrow$ x. A grammar is called left-linear (right-linear), if every rule is left-linear (right-linear) or closed. A grammar is called type-3 grammar, if every rule is right-linear or left-linear. A *Backus-System* is a type-3 grammar (Gardner *et al*., 1991). In the case of Backus-Systems each variable symbol is surrounded by brackets < and >. Moreover, there are specific syntactic symbols: ::= - the symbol for the definition process and | - the "or" symbol between sub strings.

EXPANDED BACKUS-SYSTEM

The specification of any nucleotide sequence is based on the following features: every functional unit has a specific length and many functional units are characterized by a specific frequency of sub-sequences. Therefore, it is necessary to expand the Backus-System in order to realize the formalization of these features. First of all we define an operator which will allow the usage of a specific rule for x times (k ≤ × ≤ k'). Let be i = 1..p, j = 1..q, α_i, $ß_i$, $\chi_j \in (\Phi\text{-}\{A\} \cup \Sigma)^*$ and k≤k' with k, k', p, q ε IN. The following syntactic extension expands the Backus-System:

(k:k') <A> ::= α_1 <A> $ß_1$ | ... | α_p <A> $ß_p$

<A> ::= χ_1 | ... | χ_q

The signification of this extension is defined as follows:

- <A> ::= <A≈k'> | <A≈k'-1> | ... | <A≈k>

  where <A≈i> for i = k..k' are new variable symbols and

- <A≈i> ::= α_1 <A≈ (i-1)> $ß_1$ | .... | α_p <A≈i-1> $ß_p$

  <A≈0> ::= χ_1 | ... | χ_q

where <A≈i> for i = 0..k are new variable symbols.

Notation: (k:k') is called the derivation-frequency of the rule. In the case of (k:k) we can also use the shorter description (k).

Today many different promoter units are isolated and sequenced. The analysis of this data reveals its features (Gardner *et al*., 1991). This features of the promoter sequence can be defined with the expanded Backus-System: ({Promoter, Pribnow-box, AT_P, GC_P, AT, GC, SEQ, SEQN}, {A,T,G,C}, R, Promoter)

<Promoter> ::= <GC_P> <AT_P> <SEQ> <Pribnow-box> <SEQN>

(11) <AT_P> ::= <AT> <AT_P> | <GC> <AT_P> , <AT_P> ::= A | T

(11) <GC_P> ::= <GC> <GC_P> | <AT> <GC_P>

<GC_P> ::= G | C <AT> ::= A | T <GC> ::= G | C

(5) <Pribnow-box> ::= <AT> <Pribnow-box> | <GC> <Pribnow-box>

<Pribnow-box> ::= <AT> | <GC>

(11) <SEQ> ::= <AT> <SEQ> | <GC> <SEQ>, <SEQ> ::= <AT> | <GC>

(21) <SEQN> ::= <AT> <SEQN> | <GC> <SEQN>,<SEQN> ::= <AT> | <GC>

In order to realize the second demand it is necessary to expand the definition of the derivation which will allow us to choose sub-rules. Let p,p',q,q' ε IN^+ and p/q ≤ p'/q'. The addition rule is given by:

<C> ::= <A> <C> | <B> <C> (p/q,p'/q') which means:

If we use this rule in a derivation then the probability of the application of <C> -> <A> <C> is x with $p/q \le x \le p'/q'$.

The meaning of this construct is: any word w belongs to the described language, if

1) there exists a derivation into w (in the sense of the old definition);

2) w represents a right-derivation with the following property: if we apply a rule of our new class to a derivation (<C> ::= <A> <C> | <B> <C>) then we have to consider the decomposition of the right derivation into $S \Rightarrow^* \alpha<C>x \Rightarrow^* \alpha yx \Rightarrow^* w$. αyx arises from the last application of a rule whose right side is different from C. In this case: $p/q * \#(y) \le \#_A(y) \le p'/q' * \#(y)$ and $\#_A(y)$ denotes the frequency of the appearance from A into y.

Moreover, we can also combine the described extensions of the Backus-System.

Example: representation of the promoter sequence using an extended Backus-System

({Promoter, Pribnow-Box, AT_P, GC_P, AT, GC, SEQN}, {A,T,G,C}, R, Promoter)

<Promoter> ::= <GC_P> <AT_P> <SEQ> <Pribnow-box> <SEQN>

(11) <AT_P> ::= <AT> <AT_P> | <GC> <AT_P> (6/7,1), <AT_P> ::= A | T

(11) <GC_P> ::= <GC> <GC_P> | <AT> <GC_P> (6/7,1)

<GC_P> ::= G | C <AT> ::= A | T <GC> ::= G | C

(5) <Pribnow-box> ::= <AT><Pribnow-box> | <GC><Pribnow-box> (6/7,1)

<Pribnow-box> ::= <AT> | <GC>

(11) <SEQ> ::= <AT> <SEQ> | <GC> <SEQ>, <SEQ> ::= <AT> | <GC>

(21) <SEQN> ::= <AT> <SEQN> | <GC> <SEQN>,<SEQN> ::= <AT> | <GC>

APPLICATION

The previous section shows how to describe nucleic acid using expanded Backus-Systems. The presented formalization allows the specification of DNA functional units. However, if we know a set of sequences for any DNA functional class we are able to calculate the features of this class with statistical methods. These features can be expressed using the expanded Backus-System. As an example for further analysis based on this formalization we will discuss the complexity of nucleic acid. This example belongs to the research field of sequence analysis. It has been observed that the classical complexity measures are inadequate for the purpose of characterizing biological complexity. Therefore, (Atlan, Koppel, 1990) introduced a new measure of “meaningful complexity” which they called sophistication. It was developed by modifying the classical (Kolmogoroff, 1965) program-length complexity. The program-length complexity of an object is the length of the shortest description of that object. Thus, objects which are completely characterized by some simple properties have low complexity. Objects which have no characterizing properties, and can therefore be described only by enumeration are maximally complex, and are called “random”. This definition uses the distinction between two different parts of a description of an object. The first part consists of its properties. This set of properties constitutes the object's structure. This structure might in fact be common to a whole class of objects; thus we regard a given structure as defining a class of objects. The second part of an object's description consists of the specification of the object from among the class of objects defined by its structure. Our expanded Backus-System allows the specification of DNA functional units. Furthermore, we can easily define a complexity measure:

The *value of a rule* is given by the multiplication of its frequency number (1, if there is no frequency number) and the number of its sub strings on the right side of the rule. The sum of all values of all rules which belong to the Backus-System is called *complexity of the Backus-System*.

Example: Consider a hypothetic functional unit class, which is characterized by the following sequence: an AT-rich sequence (20 bp) which is followed by the sequence

ATTA and a GC-rich sequence (10 bp). This sequence could be specified by the following expanded Backus-System: <EXAMPLE> ::= <AT> <SEQ> <GC>

(19) <AT> ::= <AT> T | <AT> A, <AT> ::= T | A

(9) <GC> ::= <GC> G | <GC>C, <GC> ::= G | C

(3) <SEQ> ::= A <SEQ> | T <SEQ> | C <SEQ> | G <SEQ>, <SEQ> ::= A | T | G | C

The complexity of this Backus-System is 71. Therefore, the complexity of the specified promoter sequence (see previous section) is 125.

REFERENCES

Atlan H., Koppel M. (1990) *Bull. of Mathematical Biology*, **52**, 335–348.
Brendel V., Busse H. (1984) *Nucl. Acids Res.*, **12**, 2561–2568.
Collado-Vides J. (1991) *J. Theor. Biology*, **148**, 401–429.
Gardner E., Simmons M., Snustad D. (1991) *Principles of Genetics*, John Wiley & Sons, New York.
Kolmogoroff A. (1965) *Prob. Inform. Transmission*, **1**, 427–467.
Maurer H. (1977) *Theoretische Grundlagen der Programmiersprachen*. BI Verlag, Mannheim.
Tanaka H. (1992) *News Generation Computing*, **10**, 329–333.

SITECON: A QUALITY TOOL FOR PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES NOW HANDLES THOSE FOR SF-1. EXPERIMENTAL VERIFICATION AND ANALYSIS OF REGULATORY REGIONS OF ORTHOLOGOUS GENES

Ignatieva E.V.*, Oshchepkov D.Yu., Klimova N.V., Vasiliev G.V., Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: eignat@bionet.nsc.ru

Key words: transcription factor binding site prediction, endocrine system, SF-1, conformational and physicochemical DNA properties

SUMMARY

Motivation: Methods that accurately predict transcription factor binding sites (TFBS) have always been important tools in studying the regulatory regions of eukaryotic genes. It is therefore important that more new high-performance methods for TFBS prediction be developed and their accuracy, assessed using experimental data.

Results: Using a new technique, SITECON, potential binding sites for the transcription factor SF-1 have been predicted in the 5′-flanking regions of a range of vertebrate steroidogenesis genes, for which it was unknown weather or not SF-1 participate in the regulation of their expression. A high predictive capacity of SITECON was proved by experimental verification: the predicted sites were all shown to be able to bind to SF-1 *in vitro*. Most of them are found at positions, which are similar to those at which known SF-1 binding sites with an experimentally proven functionality are located in the genes of other species. The new genes that we have thus detected are in fact potential targets for SF-1 and are perceived to be promising candidates for further experimental verification.

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

INTRODUCTION

Computer-based methods that predict binding sites for transcription factors are some of the most promising approaches, which it is believed can unravel the regulatory code of DNA. Statistical analysis of sample transcription factor binding sites allows their common contextual and context-dependent properties used for prediction of potential binding sites to be revealed. We have recently described SITECON (Oshchepkov, 2004a) (http://wwwmgs.bionet.nsc.ru/mgs/programs/sitecon/), our new development for determining the conservative context-dependent conformational and physicochemical properties of in transcription factor binding sites alignments. The properties so determined can be efficiently used for enhancement of binding site prediction accuracy. We have previously demonstrated how the method works on the binding sites for heterodimeric complex E2F/DP (Oshchepkov, 2004b). The discovered specific conservative properties for a set of these binding sites reflect the molecular mechanism of the heterodimer-DNA interaction.

We herein demonstrate SITECON performance on transcription factor SF-1 binding sites. The transcription factor SF-1 belongs to the family of nuclear receptors and binds to DNA as a monomer (Val *et al*., 2003). This factor plays a key role in the transcriptional regulation of steroidogenesis genes and is required for normal development of the hypothalamic-pituitary-adrenal and gonadal complexes (Busygina *et al*., 2003; Val *et al*., 2003). Experimental verification of SITECON predictions was performed and the location of the predicted sites was compared to the location of the functional SF-1 sites in the orthologous genes; the descriptions to these functional sites were taken from the literature.

METHODS AND ALGORITHMS

Nucleotide sequences of SF-1 binding sites and 5′-flanking regions. The training sample comprised the nucleotide sequences of 54 experimentally identified SF-1 binding sites retrieved from the TRRD database (Kolchanov *et al*., 2002). We were searching 5′-flanking regions of genes in two groups: a) 33 steroidogenesis genes with no experimental evidence for SF-1 binding sites in their regulatory regions; b) genes orthologous to those in first group that, according to TRRD, contain experimentally identified binding sites for SF-1.

SITECON. As the detection threshold, SITECON employs conformational similarity (Oshchepkov *et al*., 2004a), which was 94 % for SF-1. The sensitivity to type I errors was assessed using the jack-knife method: sequences were removed from the training sample one by one each in a series of iterations and served as controls. Type II errors were assessed based on the number of binding sites predicted to be present in a negative sequence 500,000 bp in length. That negative sequence was generated by random shuffling of the nucleotides of the sequences in the training sample; thus, the nucleotide compositions of both the positive and negative samples were identical and the search was made in both directions. Evaluation of type I and II is shown in Table 1.

Table 1. Errors in SF-1 binding site prediction by SITECON calculated for various conformational similarities

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

Experimental verification of the potential binding sites for SF-1. For verification purposes, a gel retardation assay of labeled 32-bp double-stranded oligonucleotide probes corresponding to the predicted binding sites was performed. The source of SF-1 was testicle cell nuclear extracts from Wistar rats aged 14 days. If the corresponding retardation bands disappeared after adding antibodies to SF-1 (Upstate), the presence of SF-1 in DNA-protein complexes was assumed.

IMPLEMENTATION AND RESULTS

Detection of new potential binding sites for SF-1 in steroidogenesis gene promoter regions with SITECON and experimental ascertainment. SITECON detected 15 new SF-1 binding sites in the promoter regions of 33 steroidogenesis genes (Table 2A). These promoter sequences had previously not been tested for binding with SF-1. Additionally, we tested three more new potential binding sites predicted in 5′-flanking gene regions, which, according to TRRD, contained experimentally identified SF-1 binding sites (Table 2B). Two of these potential binding sites were predicted to be located in the human and rat *Cyp17* genes (at positions –44 and –309, respectively).

SITECON suggests that the conformational similarity between the third potential binding site, which is at position –54 in the pig *LHbeta* gene, and the sequences of the known SF-1 binding sites is below the accepted threshold value. Because that binding site was located similarly to the known SF-1 binding sites in the orthologous (bovine, horse and rat *LHbeta*) genes, it was tested, too. All the predicted sites were tested by a gel retardation assay with antibodies. The ability to interact with SF-1 was confirmed for all the 18 binding sites (Table 2).

Table 2. Potential binding sites for SF-1 predicted with SITECON in steroidogenesis gene and experimentally ascertained

	Gene	SF-1 binding site position*	P**	Confirmed experimentally
	A			
1	*Cyp17* (Mouse)	-283	0.944	+
2	*Cyp17* (Mouse)	-49	0.949	+
3	*Ad* (Bovine)	-428	0.962	+
4	*Cyp11B1* (Guinea pig)	-126	0.945	+
5	*Cyp11B3* (Rat)	-309	0.945	+
6	*Cyp11B1* (Sheep)	-337	0.947	+
7	*Oxt* (Mouse)	-164	0.966	+
8	*Oxt* (Rat)	-167	0.962	+
9	*Oxt* (Human)	-159	0.961	+
10	*Cyp11B2* (Rat)	-324	0.951	+
11	*HSD3b* (Mouse)	–113	0.942	+
12	*Ad4BP/SF-1* (Mouse)	–224	0.952	+
13	*CYP17* (Porcine)	–51	0.946	+
14	*HSD17BI* (Rat)	–84	0.941	+
15	*LH beta* (Porcine)	–114	0.959	+
	B			
16	*CYP17* (Human)	-44	0.944	+
17	*CYP17* (Rat)	-309	0.944	+
18	*LHbeta* (Ss)	-58	0.928	+

* Position is given relative the transcription start site. **Conformational similarity to the known SF-1 binding sites as assessed by SITECON.

Analysis of transcription factor SF-1 binding site localization in the regulatory regions of orthologous genes. We compared the locations of the SF-1 binding sites in the regulatory regions of the orthologous genes using TRRD, experimental data published in the literature, and our results. The regulatory regions of five orthologous groups are presented in Fig. 1. In case of the *Cyp17*, *LHeta*, *Cyp11B1*and *Oxt* genes (Fig. 1*a*, *b*, *c*, *d*), the predicted binding sites are at the similar positions as in the orthologous genes.

DISCUSSION

It was experimentally confirmed that all the binding sites predicted with SITECON (15 at the first stage (Table 2A) and two at the second stage (Table 2B)) are able to interact with SF-1. Additionally, the –60/–50 region of the pig LHbeta gene with the conformational similarity to the known SF-1 binding sites below the threshold (0.94), too, was found to be able to interact with that transcription factor. Analysis of the regulatory regions of steroidogenesis genes suggests that the positions of most of the predicted SF-1 binding sites are similar to those of the known, experimentally identified SF-1 binding sites with proven functionality. These and our experimental data on SF-1 binding provide strong evidence that the predicted sites must be functional. The new genes that we have revealed as potential targets to SF-1 seem to be worthy of experimental verification for functionality.

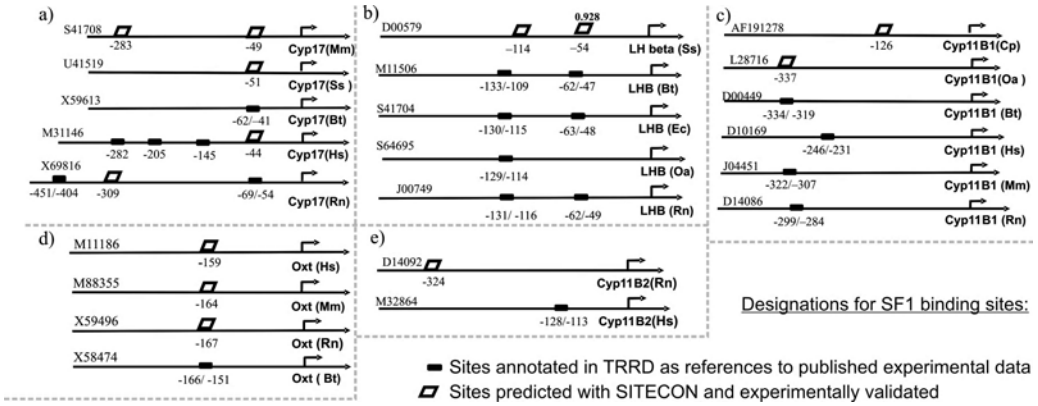


Figure 1. The regulatory regions in groups of orthologous genes for steroidogenesis. The curved arrow indicates the transcription start site. To the left above the designation of each sequence, the EMBL acc. number is indicated. Species designation: Hs, *Homo sapiens*; Mm, *Mus musculus*; Rn, *Ratus norvegicus*; Bt, *Bos taurus*; Ss, *Sus scrofa*; Ec, *Equus caballus*; Oa, *Ovis aries*; Cp, *Cavia porcellus*. 0.928 is the SITECON-based conformational similarity to the known SF-1 binding sites in the training sample.

ACKNOWLEDGEMENTS

The work was supported by Integration Project no. 119 from SB of RAS, State Contract with the Federal Agency for Science and Technology "Identification of potential targets for novel medicinal drugs based on reconstructed gene networks", Frontiers in Genetics "Living Systems", Innovation Project It-CP.5/001 "Development of software for computer modeling and design in postgenomic systems biology (*in silico* systems biology)" from the Federal Agency of Science and Innovation. The authors are grateful to V. Filonenko for translating this paper from Russian into English and to A.V. Osadchuk and T.V. Busygina for fruitful discussion.

REFERENCES

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

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

Oshchepkov D.Y., Vityaev E.E., Grigorovich D.A., Ignatieva E.A. Khlebodarova T.M. (2004a) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl. Acids Res.*, **32**, W208–W212.

Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. (2004b) SITECON—A tool for analysis of DNA physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition. In Kolchanov N., Hofestaedt R. (ed.), *Bioinformatics of Genome Regulation and Structure.* Kluwer Academic Publishers, Boston/Dordrecht/London, pp. 93–102.

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

CONTEXT-DEPENDENT EFFECTS OF UPSTREAM A-TRACTS ON PROMOTER ELECTROSTATIC PROPERTIES AND FUNCTION

Kamzolova S.G., Osypov A.A.*, Dzhelyadin T.R., Beskaravainy P.M., Sorokin A.A.
Institute of Cell Biophysics, RAS, Pushchino, Moscow region, Russia
* Corresponding author: e-mail: ao@icb.psn.ru

Key words: promoter, A-tracts, electrostatic pattern, recognition

SUMMARY

Motivation: Analysis of electrostatic properties of promoter DNA is a promising means for yielding information about promoter recognizable elements and their functioning.

Results: Electrostatic potential distribution of synthetic consensus-like promoters and their derivatives containing A-tracts at different positions in upstream region of promoter DNA was calculated and analyzed in respect with their functional behavior. Specific electrostatic motifs found in the upstream region of A-tracts containing promoters were shown to be involved as signal elements in differential recognition of the promoters by RNA polymerase α-subunit acting at early steps of complex formation.

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

INTRODUCTION

Here, electrostatic properties of three synthetic promoters P_{s1}, P_{s2}, P_{s3} and their derivatives P_{s1}/A_3-40, P_{s1}/A_3-44 and P_{s1}/A_3-48, containing 3 A-tracts at different positions in upstream region of P_{s1} were studied. All these promoters have been earlier characterized in details in comparative experiments by their interaction with RNA polymerase at different steps of complex formation and transcription initiation (Ellinger *et al*., 1994a, b). The choice of these promoters for our study was motivated by their unusual functional characteristics differing from "consensus sequence rule" behavior thus stimulating a search of new promoter determinants. The results obtained in our work indicate that electrostatic characteristics of promoter DNA can be responsible for the interaction with RNA polymerase acting at early steps of complex formation.

METHODS

Three synthetic promoters P_{s1}, P_{s2}, P_{s3} and their biochemical characterization were taken from (Ellinger *et al*., 1994a). Three derivatives of P_{s1} promoter P_{s1}/A_3-40, P_{s1}/A_3-44 and P_{s1}/A_3-48, containing three phased A-tract sequences located at different positions in upstream region were taken from (Ellinger *et al*., 1994b). Their functional characterization are presented in accordance with the data (Ellinger *et al*., 1994b).

The electrostatic potential distribution around double-helical DNA of the promoters was calculated by the Coulomb method (Kamzolova *et al*., 2000) using the computer program of Sorokin A. (lptolik@icb.psn.ru).

RESULTS AND DISCUSSION

The promoters P_{s2} and P_{s3} have consensus sequences in -10 and -35 regions, also P_{s3} has consensus 17 bp spacer between them. In P_{s2} it is 16 bp length due to 1 bp deletion of P_{s3} in -14 position. Thus, sequences of P_{s2} and P_{s3} are identical except this deletion. The P_{s1} has -35 consensus and 17 bp spacer but its -10 hexamer differ from consensus at -12. The homology scores for P_{s1}, P_{s2} and P_{s3} are 59 %, 61 % and 71 %, respectively. All the three specify the correct initiation of the expected transcript *in vivo*. But their strengths (3.4; 8.4 and 2.1 for P_{s1}, P_{s2} and P_{s3} (Ellinger *et al*., 1994a)) do not correlate with the match of the promoter sequences to the consensus pattern. P_{s3} with the highest homology score is the weakest with only one-quarter of the activity of P_{s2}. P_{s1} and P_{s2} with similar homology scores differ by a factor of 2.5 in activity. Also activities of P_{s1}, P_{s2} and P_{s3} are determined by different rate limiting steps within the pathway of RNA polymerase-promoter interaction: P_{s1} is rate-limited during early phase of the process when the enzyme binds to the promoter, P_{s2} and P_{s3} are limited in a late step involving promoter clearance in transcribing complexes.

Electrostatic profiles of P_{s2} and P_{s3}, that are limited in late steps of productive complex formation are very similar whereas P_{s1} which is rate-limited during initial binding of RNA polymerase to the promoter is characterized by quite different electrostatic pattern (Fig. 1*b*). Because electrostatic interactions contribute to promoter activity at the very early steps of RNA polymerase-promoter recognition (Kamzolova *et al*., 2000), it is reasonable to suggest that variations in functioning of P_{s1} as compared with P_{s2} and P_{s3} can be at lest partly due to the difference in their electrostatic properties. Then, in the case of P_{s1}, electrostatic component may play a role in specifying the pathway of the interaction of the promoter with RNA polymerase as well as in determining its strength. In the case of P_{s2} and P_{s3}, which are characterized by the same type of RNA polymerase-promoter interaction and very similar electrostatic patterns, some other factors can be responsible for the unpredictable difference in their activities, like different spatial arrangement of recognizable modules in the two promoters (16 bp spacing for P_{s2} and 17 bp spacing for P_{s3}) leading to overstabilization of open complexes with a lower productivity at one of them (P_{s3}) (Ellinger *et al*., 1994a).

It was shown that A-tracts inserted into upstream region of P_{s1} promoter can influence its function by increasing promoter activity due to facilitated RNA polymerase binding in the presence of A-tracts via some additional contacts between UP-region and α-subunit (Ellinger *et al*., 1994b), but mechanisms of such interaction remain unknown.

Since electrostatic properties of promoter DNA were shown to be important for the interaction with α-subunit (Kamzolova *et al*., 2000; Kamzolova *et al*., 2005) we decided to study how the insertion of A-tracts in upstream region of P_{s1} could influence its electrostatic pattern and to analyze it in respect to the functional consequences.

Three derivatives of P_{s1} containing 3 phased five-member A-tracts located at different positions in upstream region of the promoter were used (Fig. 2*a*). The first A-tract is centered around positions -40, -44 and -48 in promoter P_{s1}/A_3-40, P_{s1}/A_3-44 and P_{s1}/A_3-48, respectively. P_{s1} activity was shown to be stimulated by A-tracts in all three constructs. The strengths correspond to 3.4, 17.1, 10.6 and 10.8 for P_{s1}, P_{s1}/A_3-40, P_{s1}/A_3-44 and P_{s1}/A_3-48, respectively (Ellinger *et al*., 1994b). Maximal activation (fivefold) was observed for P_{s1} containing A-tracts at position -40. It should be noted that the stimulating effect was the same for P_{s1}/A_3-44 and P_{s1}/A_3-48 which are characterized by almost half turn dislocation of A-tracts with respect to P_{s1} core promoter sequence thus indicating no determinant role of A-tract induced DNA bending in activation of these promoters.

a

```
      -40       -30       -20       -10       +1        10        20
atgcctgcagaatgttgacagatctaccactggcggtgataattaatTgaattccggatccggcatca Ps1
atgcctgcagaatgttgacagatcttacaatgcat ctataataccacTgaattccggatccggcatca Ps2
atgcctgcagaatgttgacagatcttacaatgcatgctataataccacTgaattccggatccggcatca Ps3
```

b

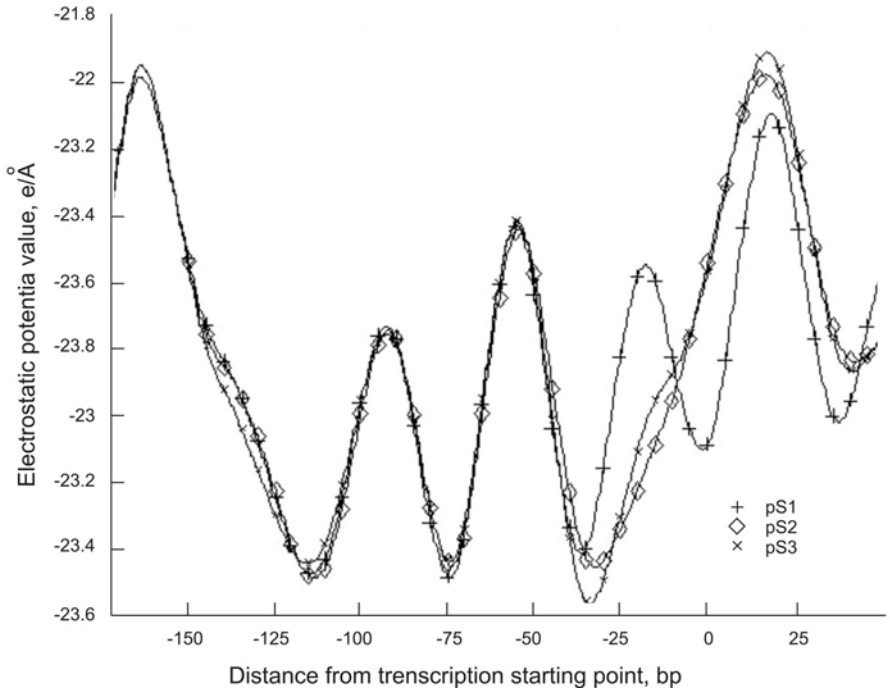


Figure 1. Electrostatic potential distribution around double-helix DNA containing Ps1, Ps2 and Ps3 promoters: *a* – nucleotide sequences of the promoters; *b* – electrostatic patterns.

The insertion of A-tracts in any position in upstream region of P_{s1} strongly influences electrostatic properties of the promoter introducing many changes in its electrostatic pattern (compare Fig. 1*b*, curve P_{s1} and Fig. 2*b*). It is noteworthy that electrostatic changes cover many sequences including those that are very far apart: the A-tracts are inserted upstream from position -40 and changes in the electrostatic profiles are observed in core sequences and downstream from the transcriptional start. The results indicate that there is no direct correlation between nucleotide sequence and its electrostatic pattern thus confirming independent character of promoter determinants based on electrostatic characteristics of promoter DNA and its structure. Electrostatic properties of DNA in far upstream region corresponding to -75 - - 100 bp positions (indicating by vertical lines in Fig. 2*b*) are of most interest for our task since this region is known to be involved in electrostatic interaction with RNA polymerase α-subunit (Kamzolova *et al.*, 2000; 2005). Fig. 2*b* shows that P_{s1}/A_3-44 and P_{s1}/A_3-48 constructs which are characterized by the same activation in response to the insertion of A-tracts, exhibit electrostatic patterns similar in design in the far upstream region. The important feature of this pattern is a continuous rise of electrostatic potential at –80 bp - -90 bp with extended positive peak in this region.

a

```
          -60        -50        -40        -30        -20        -10        +1
      gttaacaaaaactcgagaaaaatcgataaaaagcttgacagatctaccactggcggtgataatttaatTgaa Ps1/A3-40
   gttaacaaaaactcgagaaaaatcgataaaaagtcgacttgacagatctaccactggcggtgataataccacTgaa Ps1/A3-44
gttaacaaaaactcgagaaaaatcgataaaaagtcgatcgacttgacagatctaccactggcggtgataataccacTgaa Ps1/A3-48
```

b

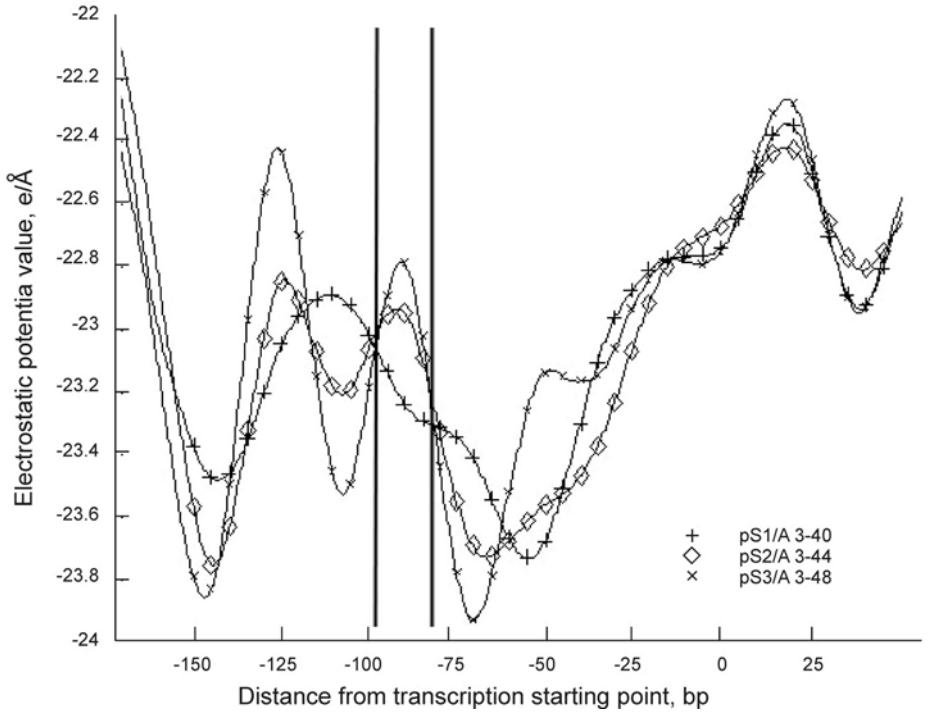


Figure 2. Electrostatic potential distribution around A-tracts containing promoters Ps1/A_3-40, Ps1/A_3-44 and Ps1/A_3-48: *a* – nucleotide sequences of the promoters; *b* – electrostatic patterns.

A distinctly different electrostatic element is found in the far upstream region of Ps1/A3-40 which is characterized by a much more stimulating effect in response to the insertion of A-tracts. Its specific feature is a more negatively charged character of –80 - -90 bp region as compared with the adjacent site located further upstream.

As shown in the cases of T4 phage promoters, the presence of different electrostatic elements in this region is essential for different type of their interaction with α-subunit thus providing a differential response in promoter functioning (Kamzolova *et al.*, 2000; 2005).

Thus, electrostatic patterns of the three A-tracts containing promoters can be specified according to the presence of some functionally important distinctive motifs which may be involved in differential recognition of the promoters by RNA polymerase α-subunit thus accounting for the difference in their functional behavior.

ACKNOWLEDGEMENTS

We are grateful to E.G. Saveljeva for technical support. The work was supported by the Russian foundation for basic research (grant RFBR-naukograd No. 04-04-97275).

REFERENCES

Ellinger T., Behnke D., Bujard H., Gralle J.D. (1994a) Stalling of *Escherichia coli* RNA polymerase in the +6 to +12 region *in vivo* is associated with tight binding to consensus promoter elements. *J. Mol. Biol*., **239**, 455–465.

Ellinger T., Behnke D., Knaus R., Bujard H., Gralle J.D. (1994b) Context-dependent effects of upstream A-tracts. *J. Mol. Biol*., **239**, 466–475.

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

Kamzolova S.G., Sorokin A.A., Dzhelyadin T.D., Beskaravainy P.M., Osypov A.A. (2005) Electrostatic potentials of *E.coli* genome DNA. *J. Biomol. Struct. Dyn.*, **23**(3), 341–346.

IDENTIFICATION OF NEW SUPEROXIDE DISMUTASE TRANSCRIPTS IN PLANTS BY EST ANALYSIS: ALTERNATIVE POLYADENYLATION AND SPLICING EVENTS

***Katyshev A.I.*[1], *Rogozin I.B.*[2], *Konstantinov Yu.M.*[*1]**
[1] Siberian Institute of Plant Physiology and Biochemistry, SB RAS, Irkutsk, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: yukon@sifibr.irk.ru

Key words: superoxide dismutase, EST analysis, alternative polyadenylation, alternative splicing, intron/exon structure, signal peptide

SUMMARY

Motivation: Superoxide dismutase (SOD) gene family in eukaryotes and prokaryotes consists of multiple genes encoding enzymes scavenging of highly toxic superoxide anion radicals. SODs are the main part of the systemic antioxidant defense against oxidative and genetic stresses. The SOD gene family plays an important role in ontogenesis of animal and plant species. A complex structural organization of antioxidant genes families (in particular, SOD gene family) is a result of numerous gene duplication events. This complex organization might reflect a high level of cell compartmentalization in plant species. The analysis of evolution of SOD genes is important for a deeper understanding of co-evolution of mitochondrial, chloroplast and nuclear genomes in plant species.

Results: In this work, the new SOD gene transcripts were identified in *Zea mays* using EST analysis. Analysis of ESTs corresponding to MnSOD gene transcripts confirmed our experimental evidence on importance of alternative polyadenylation of MnSOD gene transcripts in plant cells. The EST analysis-based identification of alternative spliced FeSOD and Cu/ZnSOD transcripts suggests that the signal peptide might be due to exon-shuffling.

INTRODUCTION

Superoxide dismutase (SOD) gene family in eukaryotes and prokaryotes consists of multiple genes encoding enzymes scavenging of highly toxic superoxide anion radicals. SODs are the main part of the systemic antioxidant defense against oxidative and genetic stresses. The SOD gene family plays an important role in ontogenesis of animal and plant species. A complex structural organization of antioxidant genes families (in particular, SOD gene family) is a result of numerous gene duplication events. This complex organization might reflect a high level of cell compartmentalization in plant species. Plants cells contain almost all known SOD types which differ by their metal cofactor and subcellular localization. Based on the metal cofactor used by the enzyme, SODs are classified into three groups: iron SOD (FeSOD), manganese SOD (MnSOD), and copper-zinc SOD (Cu/ZnSOD). FeSODs are located in the chloroplast, MnSODs in the mitochondrion and the peroxisome, and Cu/Zn SODs in the chloroplast, the cytosol, and the extracellular space.

The absence of FeSOD in animals suggested that the FeSOD gene have plastid/cyanobacterial origins and moved to the nuclear genome during eukaryotic evolution. Support of this theory comes from the existence of several conserved regions that are present in plant and cyanobacterial FeSOD sequences, but absent in non-photosynthetic bacteria. MnSOD genes, in turn, may have a mitochondrial origin. Comparison of deduced amino acid sequences from these three different types of SODs suggested that Mn- and FeSODs are more ancient types of SODs, and these enzymes most probably have arisen from the same ancestral enzyme, whereas Cu/Zn SODs have no detectable sequence similarity to Mn- and FeSODs and might be a later eukaryotic acquisition.

The most intriguing subfamily of plant SODs are FeSODs. To date, no convincing direct experimental evidence has been provided for existence of FeSOD genes in extensively studied plant species such as *Zea mays* and *Triticum aestivum*. In comparison to MnSODs, which function only in mitochondria, FeSODs and Cu/ZnSODs function in chloroplasts. It was suggested that the most ancient FeSOD enzymes had been replaced by Cu/ZnSODs in some plant species (Van Camp *et al*., 1997). Previously we have reported the identification of FeSOD gene transcript (Katyshev *et al*., 2005), as an additional chloroplast Cu/ZnSOD gene (Katyshev *et al*., 2006a) in *Z. mays*. *Z. mays* and *T. aestivum* EST analysis confirmed our experimental data and allowed us to test the aforementioned hypothesis.

We have analyzed alternatively polyadenylated and spliced forms of SOD genes which have different subcellular locations using EST analysis. Analysis of the different plant species ESTs corresponding to MnSOD gene transcripts confirmed our experimental evidence on importance of alternative polyadenylation of MnSOD gene transcripts in plant cells (Katyshev *et al*., 2006b). The EST analysis-based identification of alternative spliced FeSOD and Cu/ZnSOD transcripts suggests that the signal peptide might be acquired through exon-shuffling (Long *et al*., 1996; Vibranovski *et al*., 2006).

METHODS AND ALGORITHMS

The search of plant ESTs corresponding to SOD cDNAs was performed using the BLAST program (McGinnis, Madden, 2004) at the Plant Genome Database server (http://www.plantgdb.org/PlantGDB-cgi/blast/PlantGDBblast). We used blastn and tblastx programs with the expectation value (E-value) lower than 10^{-4}. Alignments of nucleotide sequences were constructed using a ClustalW algorithm-based program from the Vector NTI5 package (Bethesda Inc., USA). To predict subcellular localization of proteins, we used Internet resources available at the http://www.expasy.org/ molecular biology tools server: a) **the Predotar program** (Small *et al*., 2004) at the http://www.inra.fr/predotar/; b) the TargetP V1.0 program (Emanuelsson *et al*., 2000) at the http://www.cbs.dtu.dk/services/TargetP.

RESULTS AND DISCUSSION

The growing number of plant genome and transcriptome projects facilitates analysis of evolution of plant gene families. Sequencing of large genomes (e.g. *Zea mays*) is far from being complete, thus the trancriptome data analysis, such as EST analysis, is more promising. Another advantage of EST analysis is the ability to identify alternatively spliced and polyadenylated transcript variants of genes.

We have performed search of SOD cDNAs in *Z. mays* EST databases in order to delineate SOD gene composition and to provide additional support to our experimental evidence of the existence of previously undescribed FeSOD and Cu/ZnSOD genes in this plant species. Analysis of *Z. mays* ESTs corresponding to SOD gene transcripts revealed

that the real SOD gene family of this plant species is different from the earlier reported SOD repertoire (Fink, Scandalios, 2002).

Z. mays MnSOD ESTs can be divided in two large groups, corresponding to MnSOD3-1 transcript (GenBank acc. number X12540) and MnSOD3-4 cDNA (GenBank acc. number L19463). First group of sequences is further subdivided into two subgroups comprising of EST sets differing by single nucleotide substitutions. We did not find ESTs which exactly correspond to MnSOD3-2 (GenBank acc. number L19461) and MnSOD3-3 (GenBank acc. number L19462) transcripts. These results suggest that MnSOD gene family of *Z. mays* contains 2–3 gene copies similar to other monocot plant species, e.g. *Triticum aestivum* and *Oryza sativa*, and the previously reported data on the existence of four MnSOD genes in *Z. mays* genome (Fink, Scandalios, 2002) should be further revised. The data on exact MnSOD genes number in *Z. mays* could be obtained from corresponding genome regions sequencing and their mapping on chromosomes.

Both MnSOD EST groups have a substantial variability of the length of 3′-untranslated regions which could be explained by alternative polyadenylation of corresponding pre-mRNAs. Previously we reported data on alternative polyadenylation of MnSOD gene transcripts in the plant *Larix gmelinii* (Katyshev *et al*., 2006b), therefore such result is not surprising but allows us to propose possible involvement of alternative polyadenylation of MnSOD transcripts in regulation of corresponding genes expression. The identification of similar alternatively polyadenylated MnSOD transcript variants by analysis of ESTs in other plant species, *Arabidopsis thaliana* and *T. aestivum*, gave an additional support to this hypothesis.

The search for ESTs corresponding to previously reported FeSOD cDNA (Katyshev *et al*., 2005) resulted in identification of about 50 ESTs which can be divided based on sequence similarity into three major groups. The levels of sequence divergence allow us to suggest that ESTs from these three groups correspond to mRNAs of three different genes. BLAST searches for ESTs corresponding to previously described Cu/ZSOD gene transcripts (Fink, Scandalios, 2002; Katyshev *et al*., 2006a) resulted in identification of more than 500 ESTs, which can be divided based on sequence similarity into five major groups corresponding to different Cu/ZnSOD genes. These results support our experimental data and suggest that *Z. mays* SOD gene family is more complex than it was earlier suggested and contains at least four additional previously undescribed genes (Fink, Scandalios, 2002): 3 FeSOD and 1 Cu/ZnSOD genes. The accurate revision of number of MnSOD genes in *Z. mays* is also needed.

The analysis of *Z. mays* ESTs corresponding to FeSOD and Cu/ZnSOD transcripts also revealed that the large number of these transcripts has a substantial variability of exonic sequences. These results suggest that such variation of transcripts may be a consequence of alternative splicing and/or differences in intron splicing efficiency. The observed prevalence of such transcript structure alterations in 5′-terminal regions encoding N-terminal signal peptides of corresponding proteins suggests that the evolution of 5′-terminal intronic and exonic sequences of plant SOD genes may be a key mechanism of generation of duplicated SOD genes encoding enzymes of different subcellular localization.

This hypothesis is further supported by mRNA variants in other plant species: *A. thaliana* Fsd1 gene is presented by four transcript variants, Fsd1-1 – Fsd1-4, (GenBank acc. numbers NM_118642, NM_179109, NM_179110, NM_001036633). One of these mRNA variants (Fsd1-4) contains in its 5′-terminal region additional exon sequence which interrupts reading frame of the FeSOD protein, and in this case the translation from a downstream AUG codon results in the formation of active enzyme which does not contain chloroplast transit peptide. Another argument supporting the validity of this hypothesis could be obtained from analysis of the intron/exon organization of FeSOD and Cu/ZnSOD genes of *A. thaliana*, where the variation in 5'terminal intron number and size corresponds to differences in subcellular localization of the encoded proteins (Fig. 1). Based on *in silico* prediction of cellular localization of

proteins encoded by identified in EST analysis *Z. mays* FeSOD and Cu/ZnSOD genes the similar situation is characteristic and in case of *Z. mays*. Presence of exons encoding signal peptides suggests that these exons might a result of exon-shuffling, an important mechanism accounting for the origin of many new proteins in eukaryotes (Long *et al.*, 1996; Vibranovski *et al.*, 2006).

Figure 1. Intron/extron structure of FeSOD and Cu/ZnSOD of *Arabidopsis traliana*.

ACKNOWLEDGEMENTS

The work is supported by Integration projects SB RAS Nos 5.18, 6, 47.

REFERENCES

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

Fink R.C., Scandalios J.G. (2002) Molecular evolution and structure – function relationships of the superoxide dismutase gene families in angiosperms and their relationship to other eukaryotic and prokaryotic superoxide dismutases. *Arch. Biochem. Biophys*., **399**, 19–36.

Katyshev A.I., Klimenko E.S., Chernikova V.V., Kobzev V.F., Konstantinov Y.M. (2005) The new member of superoxide dismutase gene family in maize – iron superoxide dismutase. *Maize Genetics Coop. Newslett*., **79**, 38.

Katyshev A.I., Kobzev V.F., Konstantinov Y.M. (2006a) Identification of cDNA for a new chloroplast Cu/Zn superoxide dismutase in maize. *Maize Genetics Coop. Newslett*., **80**.

Katyshev A.I., Konstantinov Y.M., Kobzev V.F. (2006b) Characterization of Mn- and Cu/Zn-containing superoxide dismutase gene transcripts in *Larix gmelinii. Mol. Biol. (Mosk.)*, **40**, 372–374.

Long M., de Souza S.J., Rosenberg C., Gilbert W. (1996) Exon shuffling and the origin of the mitochondrial targeting function in plant cytochrome c1 precursor. *Proc. Natl. Acad. Sci. USA*, **93**, 7727–7731.

McGinnis S., Madden T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucl. Acids Res*., **32** (Web Server issue), W20–25.

Small I., Peeters N., Legeai F., Lurin C. (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics*, **4**, 1581–1590.

Van Camp W., Inze D., Van Montagu M. (1997) The regulation and function of tobacco superoxide dismutases. *Free Radic. Biol. Med*., **23**, 515–520.

Vibranovski M.D., Sakabe N.J., de Souza S.J. (2006) A possible role of exon-shuffling in the evolution of signal peptides of human proteins. *FEBS Lett.*, **580**, 1621–1624.

THE PREDICTION OF REGULATION OF SUBTILISIN-LIKE PROTEINASE GENE FROM *BACILLUS INTERMEDIUS* THROUGH ITS REGULATORY SEQUENCE ANALYSIS

Kayumov A.R.**, Kirillova J.M., Mikhailova E.O., Balaban N.P., Sharipova M.R.
Kazan State University, Kazan, Russia
* Corresponding author: e-mail: airat_kayumov@rambler.ru

Key words: *Bacillus intermedius*, subtilisin-like proteinase, gene expression regulation, promoter analysis

SUMMARY

Motivation: The complicated regulation of the late catabolite genes is actual problem of the modern molecular microbiology. As a model for this research serve extracellular enzymes of *Bacilli*.

Results: The nucleotide sequence of *aprBi* gene coding subtilisin-like proteinase from *Bacillus intermedius* was determined. The sequences recognized by sigma-A-RNAP and translation start site were predicted using BPROM and SignalP programs, respectively. The *aprBi* promoter analysis revealed the presence of putative sites for interaction with numerous regulatory proteins (Spo0A, DegU, AbrB, CcpA) and sigma factors. Sequences recognized by different operators and transcription sigma factors overlap each other indicating that their contributions in *aprBi* gene expression control differ in time. The participation of each transcription regulators in *aprBi* regulation was confirmed using *Bacillus subtilis* mutant strains.

Availability: Revealing of putative regulation sites in promoter region may serve as a basis for identification of regulation mechanisms that control the gene expression.

INTRODUCTION

The bacterial metabolism efficiency is provided by balance between catabolism and anabolism. Their activation and repression depends on environmental factors. Bacteria have developed mechanisms allowing coordinating metabolism in accordance with nutrients availability. The complicated regulation of the catabolite genes is reflected in their promoter architecture. Analysis of regulatory sites in promoter region allows predicting regulatory mechanisms, which control gene expression (Mironov *et al*., 1999). As a model for these researches serve microbial extracellular enzymes.

The gram-positive spore-forming bacteria *Bacillus intermedius* secrete during stationary stage of growth numerous proteinases, in which the major is subtilisin-like proteinase (Sharipova *et al*., 2002). The enzyme appears in culture liquid at the stage of slowing down of the growth, with maximal levels of the enzyme activity recorded at the 24th and 48th h of growth. Each protein fractions were isolated and characterized. The main properties of these two protein fractions were found to be similar and their N-terminal amino acid sequences appeared to be identical (Balaban *et al*., 1994, 2004). Proteinase 2 showed higher specific activity against peptide substrate. It was determined, that both enzymes are the products of one gene. However, the mechanisms involved in the

regulation of subtilisin-like proteinases synthesis during the different stages of bacterial life cycle of *B.intermedius* are still unclear.

METHODS AND ALGORITHMS

The plasmid pCS9 containing cloned *aprBi* gene was given by prof. Kostrov (IMG RAN, Moscow). The DNA fragments cloned in pCS9 were sequenced by the dideoxy chain-termination method using the T7 (Pharmacia) sequencing kit and a series of synthetic oligonucleotides that primed at intervals of approximately 300 nucleotides. Analysis of the cloned nucleotide sequence performed out using ORF Finder (Open Reading Frame Finder) network server (http://www.ncbi.nlm.nih.gov/gorf). The starting codon was detected using SignalP algorithm (http://www.cbs.dtu.dk/services/SignalP/), which allows predicting the functional activity of potential signal peptides (Bendtsen *et al*., 2004). The alignment and sequence comparisons with the GenBank database were performed with the enhanced version of the BLAST program (http://www.ncbi.nlm.nih.gov/blast) (Altschul *et al*., 1997). The DNA sequence preceding the gene for *B. intermedius* proteinase was inspected for the occurrence of the characteristic –35 and –10 boxes of SigA-type promoters (Helmann, 1995) by Softberry BPROM (Prediction of Bacterial Promoters) network server (http://www.softberry.com).

RESULTS AND DISCUSSION

The nucleotide sequence of *B. intermedius* subtilisin-like serine proteinase gene has been determined as described above and submitted to the GenBank database under accession number AY754946. The sequence analysis using the ORF Finder program revealed the presence of open reading frame coding for serine proteinase. Three putative start codons (TTG, GTG and ATG) were identified (Fig. 1). Using SignalP algorithm we have established the probability of signal peptides functional activity, starting from each of three supposed translation start sites.

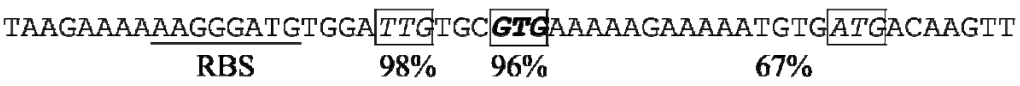


Figure 1. Putative translation start sites in *aprBi* gene. The probabilities of functional activity of corresponding signal peptides are indicated at the bottom.

Concerning analysis results, most probable are TTG (D-value = 0,79) or GTG (D-value = 0,69) (Fig. 1), not ATG (D-value = 0,23). It should be noted, that in *Bacilli* genes 10 % of ORFs are translated from GTG and 12 % start from TTG. The mutagenesis of putative start-codons has showed the translation starts from GTG.

The alignment of the *aprBi* promoter sequence with that of the gene for *B. pumilus* subtilisin-like proteinase showed 91 % identity. On the contrary, the comparative analysis of the *aprBi* and the gene for *B. subtilis* subtilisin (*aprE*) revealed only 61 % identity on extension of 81 bp in the promoter region. We propose that various regulatory pathways are involved in expression of these genes.

The *aprBi* promoter region was analyzed with respect to the putative target sequences for binding to a number of regulatory proteins. Using Softberry BPROM network server, a potential promoter sequence with poor similarity to σ^A-type –35 (score 22) and –10 (score 52) promoter recognition sequences was found in contrast with *B. subtilis aprE* gene (scores 48 and 54, respectively) (Fig. 2). It leads us to conclusion that other regulatory factors for effective *aprBi* transcription are required. The *aprBi* promoter region was examined for

putative regulatory sites. The sequences sharing 78 %, 75 % and 82 % identity with canonical sequences for interaction with σ^L, σ^H and σ^E were found (Fig. 2). The putative operator sequence for binding with carbon catabolite repressor CcpA with 78,6 % identity with canonical sequence (TGWNANCGNTNWCA) was found (Fig. 3).

Figure 2. The regulatory region of *aprBi* gene. The putative sequences recognized by transcription sigma factors are boxed.

Figure 3. The regulatory region of *aprBi* gene. A region showing homology to the consensus sequences for site binding the catabolite repressor, TGWAARCGYTWNCW and the AbrB regulatory protein, WAWWTTTWCAAAAAAW are boxed, identical nucleotides are underlined.

The *aprBi* gene expression was found to be repressing by exogenous glucose conforming its regulation by catabolite repression mechanism. Screening with WAWWTTTWCAAAAAAW, a 16-bp consensus sequence based on 20 observed AbrB binding regions, identified a region with 63 % identity (Fig. 3). The data of *aprBi* expression in AbrB-Spo0A double mutants have demonstrated the AbrB protein participation in subtilisin-like proteinase gene control. Further, in the *aprBi* gene regulatory region nucleotide sequences sharing 72–86 % identity with consensus sequence ($AGAA_{11\text{-}13}TTCAG$) typical for DegU-regulation were detected (Dartois *et al.*, 1998) (Fig. 4). These sequences appeared to be organized as direct tandem repeats. The regulatory region of *aprBi* gene contains also the sequences with structural homology (70–86 %) to specific target site for binding with Spo0A regulatory protein (TGNCGAA) (Fig. 4). Using DegS-DegU and Spo0A mutant strains was established the positive regulation of *aprBi* by these regulatory systems. Interesting, in contrast with *B.subtilis* subtilisin gene, DegS-DegU system plays minor regulatory role in *B. intermedius* subtilisin-like proteinase gene expression.

The data presented here describe the complex network regulation of *B. intermedius* serine proteinase expression, including the action of *spo0*, *degU* genes, catabolite repression and AbrB protein. This data confirm the changes of control of enzyme biosynthesis at the different stages of bacterial growth.

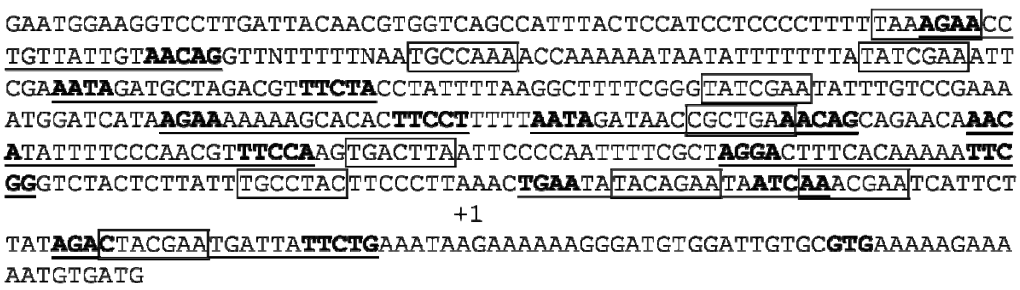


Figure 4. The *aprBi* gene promoter. Potential Spo0A binding sites are boxed. Putative DegU sites are underlined; consensus sequences for recognition by DegU are bold.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation of Basic Research (grant 05-04-48182-a).

REFERENCES

Altschul S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.

Balaban N.P. *et al*. (1994) Secreted serine protease from the spore-forming bacterium *Bacillus intermedius 3-19. Biochemistry (Moscow)*, **59**, 1033–1038.

Balaban N.P. *et al*. (2004) Isolation and characterization of serine proteinase 2 from *Bacillus intermedius 3-19*. *Biochemistry (Moscow)*, **69**, 519–526.

Bendtsen J.D. *et al*. (2004) Improved prediction of signal peptides: SignalP3.0. *Mol. Biology,* **340**, 783–795.

Dartois V *et al*. (1998) Characterization of a novel member of the DegS-DegU regulon affected by salt stress in *Bacillus subtilis. J. Bacteriol*., **180,** 1855–1861.

Helmann J.D. (1995) Compilation and analysis of *Bacillus subtilis* σ^A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucl. Acids Res.*, **23**, 2351–2360.

Mironov A.A. *et al*. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucl. Acids Res.*, **27,** 2981–2989.

Sharipova M.R. *et al*. (2002) Hydrolytic enzymes and sporulaion in *Bacillus intermedius. Microbiology (Moscow)*, **71**, 494–499.

PATTERN OF LOCALLY POSITIONED DINUCLEOTIDES IN microRNA RELATES TO ITS ACCUMULATION LEVEL

Khomicheva I.V.**, *Levitsky V.G.*, *Omelianchuk N.A.*, *Ponomarenko M.P.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: khomicheva@bionet.nsc.ru

Key words: miRNA accumulation level, genetic algorithm, discriminant analysis, locally positioned dinucleotides

SUMMARY

Motivation: microRNAs (miRNAs) are small RNA that interact with target mRNAs causing cognate mRNA degradation or translation repression, play an important regulatory role in animals and plants. Discovery of specific miRNA features in the light of experimental data on miRNA abundance allows to predict its tissue-specific expression pattern.

Results: We revealed that mutual occurrence of dinucleotides UG in positions from 17 to 19 and CA in positions from 19 to 21 (relative to 5′ end of *Arabidopsis thaliana* miRNA) corresponds to the high accumulation level of miRNAs in stems whereas the absence of both dinucleotides at the same locations corresponds to the low accumulation level. The presence of dinucleotide UG in positions from 6 to 12 together with absence of dinucleotide CC in positions from 15 to 21 corresponds to the high accumulation level of miRNAs in siliques whereas the opposite event to the low level of accumulation.

INTRODUCTION

MiRNAs are short non-coding endogenous RNA 20–24 nt long that by nearly perfect for plants (and partial for animals) complementary base pair interaction with target mRNAs lead to inhibition of translation (Olsen, Ambros, 1999; Chen, 2004) or to mRNAs cleavage (Llave *et al.*, 2002; Yekta *et al.*, 2004). MiRNA-mediated control of plant development became apparent in the comparison of miRNAs silencing deficient mutants with wild type. (Palatnik *et al*., 2003; Baulcombe, 2004; Chen, 2004). Presently in genome of *Arabidopsis thaliana* it has been found more than hundred miRNA genes (Griffiths-Jones, 2004).

Target mRNA cleavage is directed by multi-protein RISC complex (RNA-induced silencing complex, Bartel, 2004; Tang, 2005). The RISC complex consists of a dsRNAspecific RNase (DICER) along with other factors that cut the mRNA-miRNA duplex. Nucleotide context-dependent thermodynamic properties of miRNA can play a critical role in determining the molecule's function and longevity (Khvorova *et al.*, 2003). In particular, the statistical analysis of the internal stability of miRNAs precursor hairpins revealed enhanced flexibility of miRNAs precursors, especially at the 5'-anti-sense terminal region. Apparently, miRNAs have the block structure and there exists the special pattern of separately located context signals.

The miRNAs context pattern discovery is still an open problem. Known miRNAs naturally partition to the number of families of homologous sequences. Consensus based algorithms provide sequence alignment of related families only, yet common pattern discovery for sequences belonging to different families could be achieved calling for help

the secondary structure information (Griffiths-Jones *et al*., 2003). General context characteristics describing sequences of mature miRNA are not discovered yet. Existing miRNA recognition algorithms evolve all available *a priori* information about the biological model, i.e. miRNA precursor structure, complementarity to a target mRNA, conservation phenomenon of miRNA (Bengert, Dandekar, 2005).

For miRNA analysis we used SiteGA method, successfully applied earlier for transcription factor binding sites recognition (Levitsky *et al.*, in press). The highlighting advantage of SiteGA method is combination of genetic algorithm power and discriminant function insight that allows to find out subtle dependencies between local dinucleotide frequencies. In our work we revealed the dependencies between context characteristics of mature miRNA sequences found in *Arabidopsis thaliana* and microarray data on miRNA levels in the various plant organs.

METHODS AND ALGORITHMS

Mature sequences of *Arabidopsis thaliana* miRNAs were extracted from microRNA Registry database (http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/browse.pl). We used 21-nt long miRNAs as they predominantly form the miRNA pool. According to the miRNAs Registry classification these miRNAs are subdivided into 37 homologous families. Since we have merged together closely-related families of miRNAs such as: 156 and 157, 165 and 166, 170 and 171, we analyzed 34 families of sequences, totally contained 57 miRNAs.

Experimental data on miRNAs accumulation levels in various organs of *Arabidopsis thaliana* are known only for 17 miRNAs out of 57 described above (Axtell and Bartel, 2005). In total we analyzed 18 experiments for 7 plant organs: inflorescences (4 experiments), stems (2), siliques (2), cauline leaves (2), rosette leaves (2), seedlings – short days (2), seedlings – long days (2), roots (2).

To avoid the heterogeneity of source data which descended from the homology of sequences within one family and various total count of family members, we applied the iterative procedure to train the SiteGA method. Each iteratively generated train sample of sequences contained just one randomly chosen sequence from each of 34 miRNA families. In such a manner we prepared 100 TRAIN SAMPLES. The search of locally positioned dinucleotides (LPDs) was carried out by SiteGA method based on discriminant analysis and genetic algorithm (Levitsky *et al*., in press). Each LPD is characterized by the location [a,b] within miRNA [1,21] and the dinucleotide type (AA, AT…). According to SiteGA method, the positive and negative significant (Student's criterion, $p < 0.05$) correlations between LPDs frequencies for miRNAs sequences were found (100 iterations). The positive correlation of an LPDs pair denotes the most probable presence or absence of both LPDs in real sample in comparison with random (shuffled) ones. The negative correlation implies the higher probability to mutual exclusion state of LPDs pair, i.e. if the first LPD is present, then the other one is absent and vice versa. So we may conclude that both correlation types detaches two non-overlapping subsets with the contrast context characteristics. To find out the dependences between the accumulation levels of miRNAs in various plant organs and observation of LPDs absence/presence in LPDs pair the exact Fisher criterion for contingency tables was used (Table 1). Locations of dinucleotides were defined with respect to dinucleotide positions, i.e. the LPD [6;12] UG denotes that the nucleotide U should be found from the 6th to the 12th position.

Table 1. The contingency table for context feature 'negative correlation between LPD [6;12] UG and [15;20] CC' and miRNA accumulation level

Feature pattern	miRNA level, siliques (2)	
	High	Low
Presence of LPD [6;12] UG & absence of LPD [15;20] CC	3	4
Presence of LPD [15;20] CC & absence of LPD [6;12] UG	0	6

RESULTS AND DISCUSSION

The two most reliable significant dependences between LPDs correlations and miRNA levels of accumulation are shown in Table 2. The LPDs correlations corresponding to these dependences were the most frequently observed during the TRAIN SAMPLES iterations.

Table 2. Two most reliable dependences between LPDs and miRNA accumulation levels

	Organ	Dinucleotides locations and types, sign of correlation coefficient	Frequency, %	Significance, $\times 10^{-2}$	
				of dependence, Fisher's criterion	of correlation coefficient, Student's criterion
I	Stems	[17;18] UG [19;20] CC +	96	2.7	2.2
II	Siliques	[6;12] UG [15;20] CC –	38	4.9	4.1

We found significant dependences (Table 2) for only two out of seven considered tissue types (siliques, stems). The first dependence means that the mutual occurrence of dinucleotides [17;18] UG and [19;20] CA corresponds to the high accumulation level of miRNA in stems whereas the absence of both dinucleotides at the same localizations corresponds to the low accumulation level. The second dependence refers the presence of dinucleotide [6;12] UG together with the absence of dinucleotide [15;20] CC to the high accumulation level of miRNA in siliques whereas the opposite event stands for the low accumulation level.

Finally we investigated the common pattern of miRNAs features, significant LPDs correlations revealed by SiteGA method for 57 21-nt long sequences. SiteGA method allowed us to find out the most frequent significant correlations between LPDs observed for iteratively generated TRAIN SAMPLES of miRNA sequences. Five most frequently observed significant correlations ($p < 0.05$) between LPDs frequencies for the TRAIN SAMPLES are given in Table 3.

Table 3. The common miRNA features pattern: the most frequently observed significant correlations between LPDs

Significant correlations: dinucleotides locations and types[1]	Sign of correlation coefficient	Significance of correlation coefficient, $\times 10^{-2}$	Frequency, %
[4;8] CC & **[15;20] CC**	+	1.7	51
[1;1] UU & **[19;20] CA**	–	2.0	50
[17;18] UG & [17;19] UC	–	0.13	37
[19;20] CA & [19;19] CU	–	3.9	36
[6;12] UG & [13;14] CA	–	1.2	34

[1] – bold font designates LPDs involved in the significant dependences (see Table 2 above).

Note that among them we found four LPDs (Table 3, bold) that were revealed above as significantly related with miRNAs accumulation levels for siliques and stems (Table 2). The LPD [19;20] CA was found in two significant correlations, each of [15;20] CC, [6;12] UG, [17;18] UG LPD was presented in one correlation. Thus we confirmed that these LPDs are important not only for miPNAs accumulation in several tissues but they are essential miRNAs context features.

The revealed dependences allowed us to suppose that certain miRNA local context features are important for mRNA-miRNA duplex formation and stability.

ACKNOWLEDGEMENTS

This work was supported in part by Innovation project of Federal Agency of Science and innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)" No. 02.467.11.1005 of 30.09.2005, Russian Foundation for Basic Research (grants Nos 05-07-98012 and 03-04-48506), Russian Academy of Sciences (grant No. 10.4), Siberian Branch of Russian Academy of Sciences (Integration Project No. 119) and the US National Science Foundation (FIBR EF-0330786 Development Modeling and Bioinformatics), Russian Foundation for Basic Research # 05-07-90185-v, the State contract No. 02.434.11.3004 of 01.04.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction federal goal-oriented technical program "Study and development of priority directions in science and technique, 2002–2006".

REFERENCES

Axtell M.J., Bartel D.P. (2005) Antiquity of MicroRNAs and their targets in land plants. *Plant Cell*, **17(6)**, 1658–1673.

Bartel D. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Baulcombe D. (2004) RNA silencing in plants. *Nature*, **431**, 356–363.

Bengert P., Dandekar T. (2005) Current efforts in the analysis of RNAi and RNAi target genes. *Brief Bioinform*, **6(**1**)**, 72–85.

Chen X. (2004) A microRNA as a translational repressor of APETALA2 in Arabidopsis flower development. *Science*, **303**, 2022–2025.

Griffiths-Jones S. (2004) The microRNA Registry. *Nucl. Acids Res.*, **32**, 109–111.

Griffiths-Jones S. *et al*. (2003) Rfam: an RNA family database. *Nucl. Acids Res.*, **31(**1**)**, 439–441.

Khvorova A. *et al*. (2003) Functional siRNAs and miRNAs exhibit strand bias. *Cell*, **115(2)**, 209–216.

Levitsky V.G. *et al*. (2006) Method SiteGA for transcription factor binding sites recognition. *Biofizika*, **51(4)** (In Russ.) (in press).

Llave C. *et al*. (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, **14**, 1605–1619.

Olsen P., Ambros V. (1999) The lin-4 regulatory RNA controls developmental timing in *C. elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol*., **216**, 671–680.

Palatnik J.F. *et al*. (2003) Control of leaf morphogenesis by microRNAs. *Nature*, **425**, 257–263.

Tang G. (2005) siRNA and miRNA: an insight into RISCs. *Trends Biochem Sci*., **30(2)**, 106–14.

Yekta S. *et al*. (2004) MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, **304**, 594–596.

IDENTIFICATION OF *ARABIDOPSIS THALIANA* microRNAS AMONG MPSS SIGNATURES

***Khomicheva I.V.*[1], *Levitsky V.G.*[1], *Vishnevsky O.V.*[1], *Savinskaya S.A.*[2], *Omelianchuk N.A.*[*1]**

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia

[*] Corresponding author: e-mail:nadya@bionet.nsc.ru

Key words: miRNA, prediction, MPSS signatures, oligonucleotide motifs, locally positioned dinucleotides

SUMMARY

Motivation: MicroRNAs (miRNAs) are small noncoding RNAs that regulate expression of many genes through interaction with their mRNAs in plant, animal and viruses. One of the new powerful experimental approaches Massive Parallel Signature Sequencing method (MPSS) provides a set of small RNA sequences including novel miRNAs. The identification of novel miRNAs in this small RNA set needs developing of the special computer annotation software.

Results: We adapted the ARGO and SiteGA methods for MPSS data analysis. We found that integration of two methods appears to be the most reliable. Among the total MPSS pool we filtered about 2 % portion as the most probable miRNAs. Our prediction contained 93 new putative miRNA sequences forming 49 novel families, potential targets were found for these sequences in Arabidopsis transcriptome.

INTRODUCTION

Plant miRNAs are known to play an important role in gene regulation in wide range of biological processes such as plant development, organs morphogenesis, hormone response, sulfate assimilation, etc. (Bartel, 2004). Contemporary 118 miRNAs genes have been annotated in *Arabidopsis thaliana* genome (http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/browse.pl) and about 100 have been predicted. Recent estimate of miRNAs genes in the genome showed that their number could be far more than 1 % of genes and at least 20 % of genes are probably regulated by miRNAs (Xie *et al*., 2005). It includes the low expressed and tissue specific miRNAs, which are hardly to detect by common experimental methods. The computational-experimental approaches predicting *Arabidopsis thaliana* miRNAs and their targets are based on the following criteria (Adai *et al.*, 2004; Jones-Rhoades, Bartel, 2004; Lindow, Krogh, 2005; Wang *et al*., 2004):

- miRNAs genes belong to noncoding regions of genome;
- miRNAs precursors can form the stable hairpin secondary structure, containing miRNA sequence in the stem region and possessing a loop from 15 to 100 bases in length;
- miRNAs exhibit near perfect base pairing with their targets;
- mature miRNAs sequences are conserved in the *Oryza sativa* genome.

The MPSS method, which sequences hundreds of thousands of molecules per reaction, reveals the pool of small RNA in Arabidopsis seedling and inflorescence issuing the challenge to recognize new miRNAs among them (Lu *et al.*, 2005). Here, in this paper,

we used the integrated approach involving two learning paradigms to find the putative miRNAs in the MPSS small RNA set.

METHODS AND ALGORITHMS

Mature sequences of *Arabidopsis thaliana* miRNAs were extracted from microRNA Registry database (http://www.sanger.ac.uk/cgi-bin/Rfam/mirna/browse.pl). We used 21-nt long miRNA as they predominantly form the miRNA pool. According to the miRNA registry classification these miRNAs are subdivided into 37 families of homologous sequences. Since we have merged together closely-related families such as: 156 and 157, 165 and 166, 170 and 171, we have analyzed 34 families of sequences. The data under consideration contained exactly the one member from each of 34 families. This procedure allowed us to compile the representative miRNA sample and to avoid the incorrect accuracy estimation. Totally we have chosen for the training set 34 from 42 known miRNA families.

To solve the miRNAs recognition task we applied the integrated approach involving two learning paradigms, ARGO (Vishnevsky, Kolchanov, 2005) and SiteGA (Levitsky *et al.*, 2006). Briefly, ARGO is an approach for finding degenerated oligonucleotide motifs in nucleotide sequences. The SiteGA approach is based on the detection of locally positioned dinucleotides by genetic algorithm and discriminant analysis.

We used the same bootstrap iterations and thresholds setting procedures for both ARGO and SiteGA methods.

Recognition accuracy estimation was based on the standard bootstrap procedure. The full set of $M = 34$ miRNAs was randomly sampled 7 times into the new training subsets, each contained $0.8 \times M$ sequences. The ARGO and SiteGA methods trained on the basis of these subsets were applied to the rest sequences (control subsets). The random nucleotide sequences obtained by shuffling of control sequences were included in negative sequence samples. We counted the false positives (FP) and the true positives (TP) rates relying on the negative and control sequence samples correspondingly. At each TP rate we considered as the integrated recognition the success of both functions.

The dependences of FP rate vs. TP rate for each method separately and for integrated method at differing stringencies are given in Fig. 1.

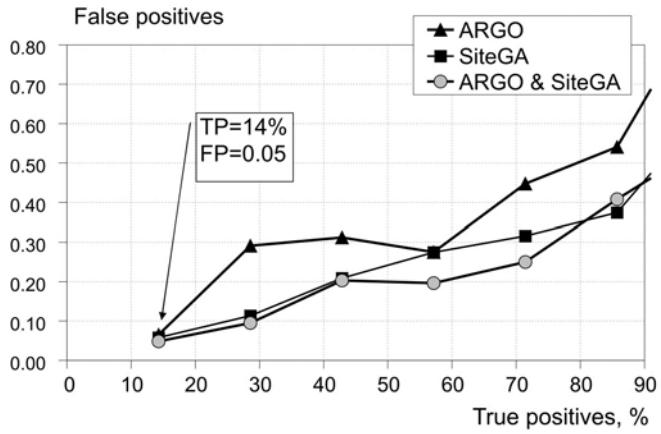


Figure 1. Recognition performance of ARGO, SiteGA methods and both taken together.

We came to the following conclusions:

- generally the recognition performance is not very high. For example, the 50% TP rate corresponds to 0.2 FP rate.

- at the most stringent threshold (TP = 14 %) both methods provide nearly the same accuracy, but at all less stringent thresholds SiteGA outperforms ARGO.
- both methods integration in the most important threshold area (TP < 60 %) allows to get higher performance against each one taken separately.

Threshold assignment procedure was based on the control data of bootstrap procedure (Fig. 1). Since accuracy estimation shows us very low prediction capacity (Fig. 1) we should apply a stringent enough threshold for analyzing real data.

We have chosen the threshold corresponding to TP = 0.14 and FP = 0.05 (Fig. 1, arrow). Then ARGO and SiteGA methods were applied the to MPSS set.

RESULTS AND DISCUSSION

The MPSS data contained nonredundant set of 33 173 signatures (Nakano *et al*., 2006). Defined above threshold allowed us to distinguish 700 distinct signatures.

The presence of the known miRNA Registry families and their members in the MPSS data set and among the ARGO&SiteGA predictions is reflected in the Table 1. MPSS data contained 78 miRNAs belonging to 22 families. These miRNAs partitioned to 18 families presented in the training set and 4 other families. We predicted exactly 28 miRNAs belonging to 15 (83 %) families from 18.

Table 1. The correspondence of known miRNAs families with the nonredundant MPSS data set and ARGO&SiteGA predictions

	No. of sequences	No. of miRNA families	No. of miRNA sequences belonging to the families
MPSS	33 173	22	78
ARGO&SiteGA prediction	700	15	28

Moreover, we came to the following:

Among the presented above ARGO&SiteGA predictions we found one new member for each of known miRNA families such as 163, 165/166, 172, and two members for each of 393 and 401 families. We revealed the rules of variability for known miRNAs within a family, variability occurs for the first two, 9th, 12th and the last three-four nucleotides in miRNAs with the conservation of other nucleotides. Only nucleotides at the permitted positions varied in the novel predicted members of known miRNA families. Essentially, that miRNA 163 was absent in the training set.

The results of integrated ARGO & SiteGA method application allowed us to predict 50 novel families totally compiling 93 miRNAs. For 34 from these 50 families we succeeded to identify the perfect complementary mRNA targets (McGinnis, Madden, 2004), which is the certain criterion for miRNAs recognition. Among these targets there were the known Arabidopsis genes, such as SPATULA, APETALA3, ETTIN, ARF2, ARF3 and ARF4, WOL (CYTOKININ RESPONSE 1), AHK4 for histidine kinase, EMF2, SPY, MYB51, CCA1 and others.

ACKNOWLEDGEMENTS

This work was supported in part by Innovation project of Federal Agency of Science and innovation IT-CP.5/001 “Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)”, Russian Foundation for Basic Research (grants Nos 05-07-98012 and 03-04-48506), Russian Academy of Sciences

(grant No. 10.4), Siberian Branch of Russian Academy of Sciences (Integration Project No. 119) and the US National Science Foundation (FIBR EF-0330786 Development Modeling and Bioinformatics), Russian Foundation for Basic Research # 05-07-90185, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction", federal goal-oriented technical program "Study and development of priority directions in science and technique, 2002–2006".

REFERENCES

Adai A. *et al.* (2004) Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res.*, **15**, 78–91.

Bartel D. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Jones-Rhoades M.W., Bartel D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, **14**, 787–799.

Levitsky V.G. *et al.* (2006) Method SiteGA for transcription factor binding sites recognition. *Biofizika* (in press).

Lindow M., Krogh A. (2005) Computational evidence for hundreds of non-conserved plant microRNAs. *BMC Genomics*, **6**, 119.

Lu C. *et al*. (2005) Elucidation of the small RNA component of the transcriptome. *Science*, **309**, 1567–1569.

McGinnis S., Madden T.L. (2004) BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucl. Acids Res.*, **32**, W20–5.

Nakano M. *et al.* (2006) Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. *Nucl. Acids Res.*, **34**, 731–735.

Vishnevsky O.V., Kolchanov N.A. (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. *Nucl. Acids Res.*, **33**, 417–422.

Wang X.-J. *et al.* (2004) Prediction and identification of *Arabidopsis thaliana* microRNAs and their mRNA targets. *Genome Biology*, **5**, 65.

Xie X. *et al.* (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.

TRANSCRIPTION FACTOR BINDING SITES RECOGNITION BY THE ExpertDiscovery SYSTEM BASED ON THE RECURSIVE COMPLEX SIGNALS

***Khomicheva I.V.*[*1, 2], *Vityaev E.E.*[2], *Shipilov T.I.*[2], *Levitsky V.G.*[1]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: khomicheva@bionet.nsc.ru

Key words: hierarchical complex signals, transcription factor binding sites, discovery

SUMMARY

Motivation: The algorithms aimed to the transcription factor binding sites (TFBSs) recognition are sensitive to context variability or to the physical-chemical, or to the conformational DNA features. The task of method development integrating the results of different recognition programs is the challenging one.

Results: We developed the ExpertDiscovery system that finds the hierarchically complicating set of complex signals. It provides the powerful tool to construct the model of regulatory region generalizing the results of different programs. Besides, the system is an independent tool to predict the TFBSs. In the paper we demonstrate that ExpertDiscovery outperforms the optimized positional weight matrix (PWM).

Availability: http://math.nsc.ru/AP/ScientificDiscovery/pages/projects.html.

INTRODUCTION

Eukaryotic regulatory regions are characterized by complex modular hierarchical structure and as the first level of organization possess the TFBSs. A pair of neighboring TFBSs organizes the so called composite element and in that case their joint action appears to be synergetic and different from if they act independently (Kel-Margoulis *et al.*, 2002). The up next level of organization consists of promoters, silencers and enhancers. The block like organization of 5′-regulatory regions means the existence of alternative promoters generally located on a considerable distance from each others. Block-hierarchical structure of eukaryotic regulatory regions provides flexible regulation on the level of transcription by switching separate elements.

Thus each level of organization states its own task in front of investigators. First of all, there is the task of TFBSs prediction, methodologically difficult by itself due to the high variety of DNA binding proteins and the tissue- and stage-specific mechanisms of regulations. The up next task is the TFBSs pattern discovery, in other words, the task of promoter localization belonging to the certain functional class according to its transcription regulation specificity (Qiu, 2003).

The problem of regulatory region analysis challenges the Data Mining and Machine Learning approaches. Machine Learning algorithms aspiring to the bioinformatics tasks are: decision trees, neural network, Hidden Markov Models, genetic algorithms, etc. (Tan, Gilbert, 2003). The traditional approach to predict TFBSs is the positional weight matrix PWM, indeed a very powerful tool, but still has some drawbacks and limitations. (Stormo, 2000).

ExpertDiscovery system (Vityaev, Shipilov, 2006), presented in the paper, finds the hierarchically complicating set of complex signals in the language of first order logic. The main advantage of ExpertDiscovery system is that it provides a powerful tool to construct the model of regulatory region generalizing the results of different programs.

METHODS AND ALGORITHMS

In "ExpertDiscovery" system the law-like rules appear to be the complex signals characterized by the set of parameters: conditional probability value, significance level (according to Fisher criterion), positive/negative coverage (the number of sequences that satisfy the complex signal). The complex signal definition is introduced recursively.

Definition 1.

1. The elementary signal (term, e.g. nucleotide, oligonucleotide) is the complex signal;
2. The result of predicates REPETITION, ORIENTATION, INTERVAL, DISTANCE implementation to the complex signal is the complex signal, i.e.:
   - REPETITION N times ($2 \le N_{\min} \le N \le N_{\max}$) of the complex signal is the complex signal. The distance between the neighbor complex signals varies in the user specified range;
   - ORIENTATION (forward, symmetric, reverse) of the complex signal is the complex signal;
   - location of the complex signal (relative to the transcription start) restricted to the certain INTERVAL is the complex signal;
   - a pair of ordered complex signals located on some DISTANCE from each other is the complex signal. DISTANCE varies in the user specified range.

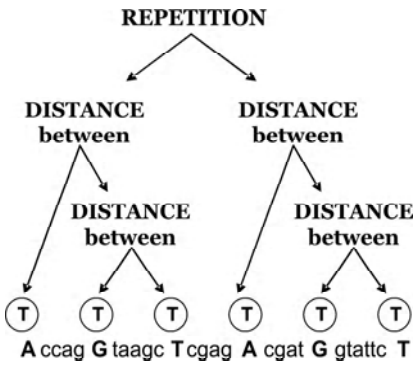


Figure 1. Complex signal hierarchical tree.The bold nucleotides-terms (denoted as "T") present how the indeed complex signal projects on the sequence. Let us follow the left branch of the tree. The nucleotide G is located from T on some distance, varying in the user-specified borders, this complex signal is located from A on the prescribed distance, and this complicated signal is REPEATED along the DNA length. The REPETITION of the complex signal is the complex signal.

According to the "Discovery" methodology ExpertDiscovery step by step complicates the current complex signals and finds all chains of nested signals. The complication is realized in a sense that the terms in the complex signal notation (Fig. 1) are replaced by the predicates REPETITION, ORIENTATION, INTERVAL, DISTANCE. The current signal becomes complicated, if the new complicated signal possesses the higher conditional probability value and the lower significance level (according to Fisher criterion). Essentially the complex signal may be expressed like the hierarchical tree (Fig. 1).

RESULTS AND DISCUSSION

ExpertDiscovery system finds the hierarchically complicating set of complex signals. The main advantage of ExpertDiscovery system is that it provides a powerful tool to formulate the verifiable hypothesis, to choose the language of prediction. First of all, you are free to organize the most suitable to the data domain list of predicates that would participate in the complex signal notion. Second, ExpertDiscovery is able to find the regularities connecting the results of different recognition programs. As the elementary signals the complex signals are based upon you can take, for example:

1. putative functional sites;
2. degenerate oligonucleotide motifs (Vishnevsky, Kolchanov, 2005);
3. sites with conservative conformational or physical-chemical features (such as double-helix angle twist, DNA melting temperature) (Oshchepkov *et al*., 2004);
4. secondary structure element (Z-DNA, RNA hairpin);
5. low complexity region (polytracks) (Orlov, Potapov, 2004).

These properties of ExpertDiscovery system provide the powerful tool to solve the complicated task – constructing the model of regulatory region generalizing the results of different programs. Moreover, the system is realized in the interactive mode with the feedback possibility, being in the dialogue with the system one can visualize the complex signal, i.e. to look through the hierarchical tree of the complex signal (Fig. 1) and to observe how the complex signal is projected to the data. The system allows to edit the complex signals, to manipulate the predicate's degrees of freedom (for example, the number of REPETITIONS, the range of INTERVAL).

As an example of ExpertDiscovery system implementation we performed the accuracy comparison of the system and the PWM according to bootstrap procedure. The train data set (sequences of SREBP BSs with flanks) was extracted from the TRRD database (Kolchanov *et al.*, 2002). Totally, the positive training set contained 38 sequences. First of all, we tried the PWM on different sequences lengths as it was described in the current issue (Levitsky *et al.*, this issue) to reach the highest PWM recognition accuracy. When the optimal sequence length for PWM was clarified and was equal to 18 nucleotides, we prepared the positive training set containing sequences of SREBP BSs of the same length. Negative training set consisted of 20 000 randomly generated sequences with the same frequencies as in the positive set (Fig. 2).

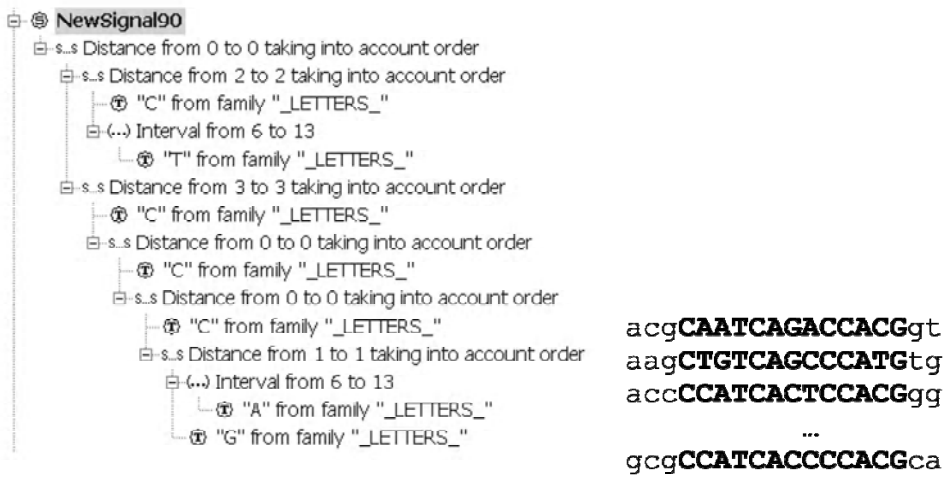


Figure 2. The hieratical tree of one of the most significant complex signals discovered for the SREBP BSs training data. Starting with the "Distance from 0 to 0 taking into account order" the tree branches to "Distance from 2 to 2 taking into account order" and "Distance from 3 to 3 taking into account order" and so on. At the left of the figure you can see how this signal is presented on the sites sequences of 18nt length (bold, capital).

The positive training set was randomly sampled 7 times into the new subsets each contained 32 sequences. The PWM and ExpertDiscovery methods trained on the basis of these subsets were applied to the rest sequences (control subsets). We counted the false positives (FP) and the false negative (FN) rates relying on the negative and control sequence samples correspondingly (Table 1). The score of the sequence was equal to the negative sum of the significance levels of the regularities the sequence satisfies to.

Table 1. False positive rates for PWM and ExpertDiscovery provided the same FN rates

FN rate	FP rate PWM	FP rate ExpertDiscovery
58%	7.4E-04	2.0E-04
54%	7.9E-04	3.5E-04
50%	8.2E-04	7.0E-04
46%	8.6E-04	8.2E-04
42%	9.9E-04	1.0E-03
38%	1.8E-03	1.3E-03

ACKNOWLEDGEMENTS

The work is partially supported by the Russian Foundation for Basic Research # 05-07-90185-v, Scientific Schools grant at the President of the Russian Federation No. 4413.2006.1, Innovation project IT-CP.5/001 “Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)”.

REFERENCES

Kel-Margoulis O. *et al*. (2002) Transcompel. *Nucl. Acids Res*., **30**, 332–334.

Kolchanov N.A. *et al.* (2002) Transcription Regulatory Regions Database, (TRRD): its status in 2002. *Nucl. Acid Res.*, **30**, 312–317.

Levitsky V.G. *et al.* (2006) The SiteGA and PWM methods application for transcription factor binding sites recognition in EPD promoters. *This issue*.

Orlov Y.L., Potapov V.N. (2004) Complexity: an internet resource for analysis of DNA sequence complexity, *Nucl. Acids Res*., **32** (Web Server issue), 628–633.

Oshchepkov D.Y. *et al.* (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition, *Nucl. Acids Res.*, **32** (Web Server issue), 208–212.

Qiu P. (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network, *Biochem Biophys Res Commun*., **309(**3**)**, 495–501.

Stormo G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Tan A.C., Gilbert D. (2003) An empirical comparison of supervised machine learning techniques in bioinformatics. *Proceedings of First Asia Pacific Bioinformatics Conference (APBC).*

Vishnevsky O.V., Kolchanov N.A. (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. *Nucl. Acids Res*., **33**, 417–422.

Vityaev E.E., Shipilov T.I. (2006) Software for analysis of gene regulatory sequences by knowledge discovery methods. In Kolchanov N., Hofestaedt R., (eds), *Bioinformatics of Genome Regulation and Structure II.* Springer Science+Business Media, Inc., pp. 491–498.

TRRD: A DATABASE ON EXPERIMENTALLY IDENTIFIED TRANSCRIPTION REGULATORY REGIONS AND TRANSCRIPTION FACTOR BINDING SITES

Kolchanov N.A.*[*], *Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Khlebodarova T.M., Merkulov V.M., Merkulova T.I., Podkolodny N.L., Romashenko A.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: kol@bionet.nsc.ru

Key words: transcription regulation, regulatory region, transcription factor binding site, gene expression

SUMMARY

Motivation: The main goal of TRRD (Transcription Regulatory Regions Database) development is the most complete and adequate description of the structural and functional organization of transcription regulatory gene regions in eukaryotes based on the data obtained experimentally.

Results: The overall information contained in the current TRRD release is represented as eight libraries: TRRDGENES, TRRDUNITS, TRRDEXP, TRRDSITES, TRRDFACTORS, TRRDLCR, TRRDSTARTS, and TRRDBIB. TRRD compiles the data on 2344 genes, 14 407 patterns of their expression, 3490 regulatory units, and 10 135 transcription factor binding sites associated with them. This database contains only experimentally confirmed information. TRRD is filled in by manual annotation of scientific publications. The data incorporated into TRRD is a result of annotation of 7609 scientific papers. The main tool for searching TRRD and navigation in it is SRS. A large number of indexed fields in the SRS version of TRRD allow the user to generate complex queries both within individual libraries and involving several libraries. TRRD has thesauruses that provide additional options for data access. The number of databases linked to TRRD has been increased.

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

INTRODUCTION

The structure–function organization of regulatory regions in the genes transcribed by RNA polymerase II is typically very intricate. The presence of alternative promoters and remote regulatory regions localized to both the 5′- and 3′-gene–flanking regions as well as to introns and exons are typical of the numerous genes studied so far. Active contributors to combinatorial gene regulation are the structural elements of core promoters (Smale, Kadonaga, 2003). Transcription factor binding sites within a regulatory unit (promoter, enhancer, or silencer) may be organized in functional modules that determine one or another expression pattern of a gene. One more functionally important characteristic is the multiple transcription starts. This particular information may be very important, as individual transcription starts of one promoter are frequently used for producing

transcripts in different tissues or under different conditions (under the action of inducers, at various ontogenetic stages, etc.).

All these facts clearly indicate that the description of an integrated system of transcription regulation requires the comprehensive information about the regulatory elements of the gene. Creation of collections of experimentally discovered data on the regulatory elements acting at all levels is absolutely necessary for both forming the concepts of what is the nature of regulation of individual genes and developing the computer methods for prediction of regulatory elements, construction of gene networks, and functional genome annotation. TRRD, which we are presenting here, is a unique information resource that is developed aiming to provide an integrated description of transcription regulation of the eukaryotic genes transcribed by RNA pol II. The database is being constantly supplemented with new information, and the TRRD format is being permanently developed. Based on the information contained in TRRD, tools for analyzing regulatory regions of the genes transcribed by RNA pol II were developed.

Structure of the TRRD database and data source

All the information contained in TRRD[2] is distributed between eight interconnected libraries. The TRRDGENES is the central, integrating library, which compiles the information identifying the gene, internal references to other TRRD libraries, and references to external databases and resources as well as hierarchically organized representation of the regulatory elements of all levels. The rest information tables of TRRD are TRRDSITES collating the information about transcription factor binding sites; TRRDUNITS describing regulatory units (promoters, enhancers, and silencers); TRRDLCR containing the structure–function characteristics the locus control regions (LCR); TRRDSTARTS containing the data on transcription initiation starts; TRRDEXP compiling the description of the qualitative specific features of gene expression; and TRRDBIB containing the bibliographic information. The description of information fields was given in detail previously (Kolchanov *et al*., 2000, 2002). TRRD is filled in by manual annotation of scientific papers. The database contains only experimentally confirmed information obtained in experiments of various types (http://srs6.bionet.nsc.ru/srs6bin/cgi-bin/wgetz?-page+FieldInfo+-lib+TRRDSITES4+-bf+ExperimentCodes). The data input is standardized via the system of controlled vocabularies.

RECENT DEVELOPMENTS

Development of the TRRD 7.0 format

The format of TRRD is being constantly developed to enhance the search of the database and simplify the data access. In TRRD release 7.0, the TRRDGENES library contains a considerably larger number of links to external databases: in addition to the previously available references to SWISS-PROT and EMBL/GenBank, note the links of the current release to Entrez Gene, GeneCards, MGI, RGD, FlyBase, and MaizeDB (overall, more than 20 databases).

A new library, TRRDSTARTS, was developed. This library compiles the data on the experimentally determined transcription start sites of genes. TRRDSTARTS contains the absolute genome coordinates of the major and minor transcription starts of genes (with indication of the chromosome and the release of genomic database).

The format of TRRDSITES library was extended. A new field, PreferredName (NP), was added; this field contains the standard (preferred) site name. The field PreferredName

[2] In the public version of TRRD, a number of information fields are not available, in particular, the sequences of transcription factor binding sites (TFBS) and regulatory regions, their localization in the corresponding entries of EMBL or GenBank database of nucleotide sequences, and the TFBS localization relative to a particular reference point within the gene.

is filled in automatically based on the data from the field TF of the block FACTOR, connected to this site. In this process, a specialized vocabulary of transcription factors is used, where the relations (the first order hierarchy and synonymy) between the names are fixed too. Thus, a query to the field PreferredName gives the possibility to obtain the entire information contained in TRRD that is related to the transcription factor binding sites of the user-specified type independently of what synonymic factor names were used in the query.

The TRRDFACTORS library was revised. The vocabularies of transcription factors were unified to assign a unique identifier to each factor. Description of the subunit composition for multimeric factors is provided.

The extension of TRRD content

TRRD is filled up constantly with the new information. The number of entries in TRRD release 6.0 (Kolchanov *et al*., 2002) and the current release 7.0 (as of September 01, 2005) are listed in Table 1.

Table 1. The information content of TRRD

Library name	Number of entries in release 6.0	Number of entries in the current release 7.0	Including the species (%)			
			Human	Mouse	Rat	Others
TRRDGENES	1167	2344	32	22	15	31
TRRDUNITS	1714	3490	36	19	14	31
TRRDEXP	5335	14 407	37	37	18	18
TRRDSITES	5537	10 135	36	18	14	32
TRRDBIB	3898	7609	37	21	16	26

The sections on a number of subjects are being developed in TRRD that include genes united according to various functional characteristics. Each of the sections contains a group of genes expressed under certain conditions or involved in a certain process. Overall, nine sections were described earlier (Kolchanov *et al*., 2002). At present, TRRD contains 18 sections of this type (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/sections1.shtml). These sections are also a tool for quick access to the information contained in TRRD.

New possibilities for access to TRRD data

SRS (Sequence Retrieval System) version 6.1.3.11, which provides searching for information over 132 indexed fields, is the main tool for accessing TRRD. In addition, several specialized search systems were developed. These systems are based on the controlled vocabularies of tissues, cells, organs, developmental stages, external stimuli, and transcription factors, on the one hand, and thesauruses on organs and tissues in mammals (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/), on the other. During operation of these searching systems, the relations of the types "general–particular", "part–whole", "synonymy", etc., are realized. The queries to the SRS version of TRRD are realized not only according to a specified term, but also by all the related terms (daughter with reference to the query term) in the corresponding vocabulary as well as by all the synonyms simultaneously. These searching systems (http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/search.html and http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/thesaurus/search_hidden.html) provide (1) search for the genes induced (or repressed) by an external stimulus; (2) search for the genes expressed in a specified organ, tissue, cell type, or stage of organism development; (3) a combined search for the genes expressed in a specified tissue, organ, or cell type when induced by a specified external stimulus (simultaneously); and (4) search for the genes or sites regulated by a specified transcription factor.

New tools for analysis of DNA sequences using TRRD

Three new tools for prediction of transcription factor binding sites and promoters were developed based on the information collected in TRRD: (1) SITECON (Oshchepkov *et al*., 2004) and SiteGA (Levitsky *et al*., 2006) for site recognition and (2) ARGO

(Vishnevsky, Kolchanov, 2005) for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters.

ACKNOWLEDGEMENTS

The authors are grateful to I.V. Lokhova for bibliographical support and to G.B. Chirikova for translation of the paper into English. The work was supported by RFBR (grants Nos 05-07-98012 and 05-04-49111), Siberian Branch of the Russian Academy of Sciences (integration project No. 119), the government contract with the Federal Agency for Science and Technology "Identification of potential targets for novel medicinal drugs based on reconstructed gene networks", the priority direction "Living systems", innovation project of Federal Agency of Science and Innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)", NATO (grant No. PDD(CP)-(LST.CLG 979815), and INTAS (project No. 2382).

REFERENCES

Kolchanov N.A. *et al*. (2000) Transcription regulatory regions database (TRRD): its status in 2000. *Nucl. Acids Res*., **28**, 298–301.

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

Levitsky V.G. *et al*. (2006) Method SiteGA for recognition of transcription factor binding sites. *Biofizika* (in Russ.) (in press).

Smale S.T., Kadonaga J.T. (2003) The RNA polymerase II core promoter. *Annu. Rev. Biochem.*, **72**, 449–479.

Oshchepkov D.Y. *et al*. (2004) SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition. *Nucl. Acids Res*., **32**, W208–W212.

Vishnevsky O.V, Kolchanov N.A. (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. *Nucl. Acids Res.,* **33**, W417–W422.

METHODS FOR RECOGNITION OF INTERFERON-INDUCIBLE SITES, PROMOTERS, AND ENHANCERS

Kondrakhin Yu.V.**, Ananko E.A., Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: eananko@bionet.nsc.ru

Key words: recognition of transcription factor binding sites, interferon-stimulated genes, genome annotation

SUMMARY

Motivation: Combination of computer-assisted genome annotation with the large-scale experimental research is helpful for determination of the functions of genes, which are not studied yet. The task of searching for target genes of interferon (IFN)[3] induction is of intense interest. IFNs modulate the operating of immune system, they exert antiviral, antibacterial and antitumoral effect. An important impact in functioning of interferon system is produced by transcription factors ISGF3, STAT1, and IRF1.

Results: By analyzing localization of binding sites of 20 various transcription factors in regulatory regions of genes referring to different functional groups, it was estimated that the regions from -500 to transcription start site of IFN-inducible genes are enriched by the binding sites for transcription factors ISGF3, STAT1, and IRF1. We have developed the methods of recognition of these transcription factor binding sites, as well as the methods of recognition of IFN-inducible promoters and enhancers.

INTRODUCTION

Interferons refer to the class of cytokines. They possess by a wide spectrum of biological activities. For example, type I IFNs, in particular, IFNs-α and IFN-β, exert antiviral, antiproliferative, and antitumoral effect: they activate the cells of immune system and modulate cell differentiation (Pestka *et al*., 2004). The type II IFN, IFN-γ, makes great impact into development of antibacterial and antiparasitic responses (Decker *et al.*, 2002). IFN-γ was also shown to participate in development of autoimmune state (Baccala *et al.*, 2005). When IFNs interact with the cell surface receptors, they activate the JAK-STAT signal transduction pathway. As a result of this process, ISGF3 and STAT1 transcription factors are activated by type I IFNs and IFN-γ, respectively (Kalvakolanu, 2003; Uddin, Platanias, 2004). Transcription factors referring to the family of IRF (Interferon Regulatory Factors) (Mamane *et al*., 1999) also play an important role in functioning of the IFN system (Barnes *et al*., 2002). Interaction of ISGF3, STAT1, as well as of some IRFs with the binding sites in regulatory regions of IFN-stimulated genes (ISG) causes enhancement of transcription of these genes.

[3] The abbreviations used are: IFN, interferon; ISG, Interferon-Stimulated Genes; ISGF3, Interferon-Stimulated Gene Factor 3; IRF, Interferon Regulatory Factor; STAT, Signal Transducer and Activator of Transcription.

We have developed the methods aimed at recognition of ISGF3, STAT1, and IRF1 binding sites that are of considerable importance for the functioning of the IFN system. By analysis of putative binding sites of 20 different transcription factors, we have revealed regularities in localization of sites in promoter regions of ISG. Based on these regularities, we have designed the methods for recognition of IFN-inducible promoters and enhancers.

METHODS AND ALGORITHMS

For recognition of transcription factor binding sites that are of primary importance for the functioning of the IFN system, in particular, ISGF3, IRF1, STAT1, NF-κB, and AP-1, we have mainly applied the matrix method based on the additive, or multiplicative recognition function. In order to obtain the frequency and weight matrices, we have developed the special iterative approach that was applied to the different samples of binding sites extracted from the TRRD database (Kolchanov *et al*., 2002). The size of the most samples was varying in the range from 30 to 70 sequences.

TO determine the weight matrix, we have used three methods of the multiple alignment, which are described by a single generalized algorithm within the frames of the Gibbs sampler approach. The algorithm of the methods suggested is iterative, so that each iteration consists of two steps. For operating of this algorithm, we need the initial approximation for the frequency matrix $F=(f_{ij})$, $i=\{A,C,G,T\}$, $j=1,\ldots,l$, where l denotes the length of a site and f_{ij} is the frequency of the occurrence of the nucleotide i at the j-th position of the aligned sequence j. At the first step of iteration, the frequency matrix F is rearranged by means of transformation T into the weight matrix W=(wij): W=T(F). By moving along the first sequence of the sample with the step of 1 bp, we calculate the value of recognition function G. For generating three methods of alignment, we have used the following variants of recognition function G and rearrangement T of the frequency matrices into the weight matrices:

1st method.

$$T: w_{ij} = f_{ij} / (f_{Aj} + f_{Cj} + f_{Gj} + f_{Tj}), \quad (1)$$

For the nucleotide sequence, $S=s_1,\ldots,s_l$, with the length l, the recognition score is calculated by using the additive function

$$G(s_1,\ldots,s_l) = \sum_{i=1,\ldots,l} w_{s_i^i}. \quad (2)$$

2nd method. We use the same rearrangement T as in the first method. The function G for the sequence S is calculated in accordance with multiplicative function

$$G(s_1,\ldots,s_l) = \prod_{i=1,\ldots,l} w_{s_i^i}.$$

3rd method. The process of generating the weight matrix consists of two stages. First, we use transformation indicated in description of the first method. Then for each position j, $j = 1,\ldots,l$ we calculate the entropy Ej by the formula

$$E_j = - \sum_{i=A,C,G,T} w_{ij} \, r \ln (w_{ij}). \quad (3)$$

The final weights wij* are obtained by renormalization of initial weights wij by the formula

$$w_{ij}^* = w_{ij} / E_j^*, \quad (4)$$

where E_j^* is a modified entropy of the j-th position, that is, $E_j^* = \{E_j, \ if \ E_j>0.1; \ \ 0.1 \ otherwise\}$.

The modified entropy is introduced due to necessity to avoid the formal division by zero in the formula (3) under the treatment of completely conservative site positions. For calculation of the score, we use the same additive function (2) as in the first method.

The threshold values of the 1st type errors were estimated with help of training samples by the jack-knife method. The threshold values of the 2nd type errors were estimated with help of control samples taken from GenBank.

RESULTS

In Table 1, the threshold values of the 1st and 2nd type errors are given (α_1 and α_2,, respectively) for the methods for recognition of transcription factor binding sites important for the functioning of ISG. In the last column of the Table 1, the level of false-negatives in the control samples is presented. In this case (last column), the control samples were compiled from the sequences of natural sites from the TRRD database, which were not included into the training samples.

Table 1. Threshold values for recognition of binding sites for transcription factors AP-1, IRF1, ISGF3, NF-κB, and STAT1

Transcription factor binding site	First type error (α_1)	Second type error (α_2)	Independent control (false-negatives)
AP-1	37 %	2.81E-03	no data
IRF1	24 %	9.59E-05	31.8 %
ISGF3	25 %	6.84E-04	46.2 %
NF-κB	42 %	5.32E-04	70.8 %
STAT1	43 %	8.82E-05	84.6 %

Comparison of recognition AP-1, IRF1, ISGF3, NF-κB, and STAT1 binding sites by the method suggested and by some other methods is given in the supplementary material (http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/BGRS_2006/supplementary.htm).

In order to find out specific organization of ISG regulatory regions, we have performed comparative analysis of three functional groups of genes. Except ISG, we have taken in analysis glucocorticoid-regulated genes (GR, 39 genes, the set is compiled by T.I. Merkulova) and lipid metabolism genes (LM, 56 genes, the sample is compiled by E.V. Ignatieva). In the supplementary material (http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/BGRS_2006/supplementary.htm), some regularities are illustrated, which differ at most between ISG and "accidentally chosen" human genes extracted as a control sample from the EPD database and the genes of the other functional groups.

While designing the recognition method for IFN-inducible genome regions, we have performed the estimation of occurrence of different site combinations. As a combination, we consider the simultaneous presence of two or three sites localized at a given distance from each other and/or in the region pre-ordered relatively transcription start. The presence of the definite types of sites at the distance given without associating them with position of transcription start has enabled us to reveal IFN-inducible enhancers and, possibly, alternative promoters of genes, for which localization of transcription start was not determined.

In total, we have analyzed several hundred of site combinations, out of which we have selected 158 sites statistically differing by occurrence from that of the control samples. By using these combinations and information about the type and induction level of each gene from the control sample, we have designed three methods aimed at recognition of IFN-inducible regions of genes:

method 0 – recognition of any IFN-inducible DNA region (stimulation by any IFN);

method 1 – recognition of DNA regions induced by type I IFNs (IFN-α, IFN-β);

method 2 – recognition of DNA regions stimulated by type II IFN (IFN-γ).

The methods were independent on the training samples, because the exclusion of training sites from the results didn't change the character of site distribution (data not shown). Details for each method are given at http://wwwmgs.bionet.nsc.ru/mgs/papers/ananko/BGRS_2006/supplementary.htm.

DISCUSSION

With the help of the methods developed, we have studied the DNA sequences from -1000 to +1000 bp relatively transcription start site of genes referring to different functional groups under various threshold values of the methods. The best recognition values were obtained by applying the method 0, which is applicable for recognition of the gene regions regulated by any IFN. Under the threshold value of the function equaling to 0.4, we have recognized 36 % of genes from the learning sample; and from 5 to 12 % of genes from the other gene samples. If the threshold value of the function was increased up to 0.7, the recognition in the learning sample falls down to 12.5 %, whereas in the rest samples, it was equaling to at most 1 %.

Studying of the human genome by the method 2 has enabled to detect 90 genes under the threshold value equaling to 0.65. These genes compiling 1 % out of the sample studied are induced by IFN-γ with 85 % probability. If the threshold value of the function is decreased to 0.3, then the recognition accuracy has grown up to 12.3 % (in total, 1023 genes). The similar studies developing the method for recognition of genes stimulated by IFN-γ were reported in 2004 (Liu *et al.*, 2004). In accordance with the estimates made by the authors, 65 % of genes predicted by this method are really induced by IFN-γ.

Application of the methods developed for computer genome annotation in combination with microarray analysis will be helpful in determining the functions of genes, which are not studied yet.

ACKNOWLEDGEMENTS

The authors are grateful to E.V. Ignatieva for kindly providing the sample of lipid metabolism genes and to G.V. Orlova for translating the manuscript from Russian into English. The work was supported by the Russian Government (Contracts Nos 02.434.11.3004, 02.467.11.1005) and Siberian Branch of the Russian Academy of Sciences (the project “Computational simulation and experimental design of gene networks” and integration project No. 115).

REFERENCES

Baccala R. *et al*. (2005) Interferons as pathogenic effectors in autoimmunity. *Immunol. Rev*., **204**, 9–26.
Barnes B., Lubyova B., Pitha P.M. (2002) On the role of IRF in host defense. *J. Interferon Cytokine Res.*, **22**, 59–71.
Decker T. *et al.* (2002) IFNs and STATs in innate immunity to microorganisms. *J. Clin. Invest.*, **109**, 1271–1277.
Kalvakolanu D.V. (2003) Alternate interferon signaling pathways. *Pharmacol. Ther.*, **100**, 1–29.
Kolchanov N.A. *et al*. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. *Nucl. Acids Res*., **30**, 312–317.
Liu H. *et al.* (2004) Prediction of IFN-gamma regulated gene transcription. *In Silico Biol.*, **4**, 489–505.
Mamane Y. *et al.* (1999) Interferon regulatory factors: the next generation. *Gene*, **237**, 1–14.
Pestka S. *et al.* (2004) Interferons, interferon-like cytokines, and their receptors. *Immunol. Rev.*, **202**, 8–32.
Uddin S., Platanias L.C. (2004) Mechanisms of type-I interferon signal transduction. *J. Biochem. Mol. Biol.*, **37**, 635–641.

GENOME-WIDE CO-EXPRESSION PATTERNS OF HUMAN CIS-ANTISENSE GENE PAIRS

Kuznetsov V.A.*, Zhou J.T., George J., Orlov Yu.L.
Genome Institute of Singapore, 138672, Singapore
* Corresponding author: e-mail: kuznetsov@gis.a-star.edu.sg

Key words: transcription, expression regulation, sense-antisense genes, microarray

SUMMARY

Motivation: To study possible mechanisms of cis-antisense regulation in human genome, we carried out a global microarray-based co-expression analysis of a large number of cis-antisense gene pairs. We focused on cis-antisense expression patterns in distinct breast cancer types.

Results: We identified common positive co-regulation patterns which reproducibly expressed in the human breast cancer cells and specific positive co-expression patterns which uniquely associated with low- and in highly-aggressive types of breast cancer cells. The enrichment of co-regulation patterns of gene pairs in the same loci in compare to random gene pairs in different loci of the human genome and absence of other regulatory modalities allowed us to suggest that cis-antisense transcripts might be controlled by (i) distant mechanisms associated with chromatin remodeling and by (ii) local mechanisms due to mutual local de-repression of mRNA synthesis initiated by temporal triplex-forming mRNA which can form DNA transcription bubbles.

INTRODUCTION

Pairs of genes transcribed from opposite strand of the same locus with antiparallel exon-exon overlaps are commonly referred to as cis-antisense pairs. It is a widely-accepted paradigm that genes in a cis-antisense pair may regulate each other, pre- or post-transcriptionally. Translational down-regulation of a sense transcript by antisense RNA induction has been observed (Vanhee-Brossollet, Vaquero, 1998; Chau *et al*., 2002). It is assumed that hybridization of two RNAs cis-antisense to one another results in translation blockage via steric hindrance and/or RNAase-mediated degradation of the duplex (Vanhee-Brossollet, Vaquero, 1998). At the transcriptional level, the two members of the cis-antisense pair can compete for transcription from the same locus (Chau *et al*., 2002). Alternatively, these genes may be controlled by other (e.g. epigenetic) mechanisms. In an attempt to distinguish between these regulatory scenarios, we carried out global gene co-expression analysis of a large number of cis-antisense pairs based on microarray data from different human cell types. We focused on regulatory patterns in human breast cancer.

Since cis-antisense pairs affect up to 25 % of genes in mammalian genomes, understanding their regulatory significance is a biological imperative. Previous studies have emphasized post-transcriptional antisense-mediated repression scenarios in both prokaryotes and eukaryotes, characterized by repression of generally protein-coding sense transcripts by their cis-antisense counterparts. This paradigm was challenged by the international FANTOM3 consortium (Katayama *et al*., 2005), which established that co-regulation (i.e. a situation where the two members of an sense-antisense pair are either both highly expressed or both repressed) was the dominant scenario for a small and biased

sample of cis-antisense pairs profiled in cell line system perturbation experiments under specific condition. We aimed to resolve the repression /co-regulation controversy, since it is relevant to pre-transcriptional (e.g. chromatin remodeling) vs. post-transcriptional (e.g. ncRNA-mediated degradation of sense via RNA duplex formation) mechanisms of cis-antisense transcriptional control.

MATERIALS AND METHODS

To link sense-antisence transcript pairs with reliable expression data we revisited chromosome coordinates of Affymentrix GeneChip U133 probesets in the human genome. Well-defined U133 (A and B) Affymetrix probesets have been selected using BLAT alignment of ~44,500 original Affimetrix sequences on the human genome. We have removed 2500 multiple mapped and other erroneous probesets (Orlov *et al*., this issue). Through genomic alignments of EST and cDNA sequences and after manual verification of transcript orientation, splice sites, and polyadenylation signals 4511 cis-antisense transcript (SAT) pairs have been selected (done by Dr. L. Lipovich's group at GIS). Then we identified 2,816 human cis-antisense transcript pairs matched by the filtered 4,458 Affymetrix probesets. The probesets pairs represent 1450 cis-antisense transcript pairs (72 % mRNA, 19 % spliced EST pairs, 9 % RefSeq).

We stored information about Affy probeset sequences matched to cis-antisense gene/mRNA/EST pairs sequences, probesets orientation, their map on chromosome, genome characteristics of the sense-antisense gene pairs, and annotation information (RefSeq, mRNA, EST, chromosome coordinates) in our local SAT database.

We define convergent cis-antisense pairs as gene pairs whose ends overlap but starts does not (tail-to-tail); divergent pairs as gene pairs whose starts overlap but ends does not (head-to-head), and complex pairs as those of any other configuration (Fig. 1). The numbers of gene pairs that expression was measured in microarray experiments is noted below corresponding schemes. The tail-to-tail topological type was most common type of pairs in our dataset (Fig. 1).

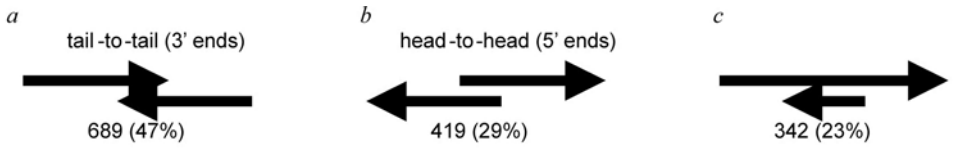


Figure 1. Three types of exon-to-exon cis-antisense pairs.

Tumor samples were derived from large cohorts (in total 251 patients) from primary human breast cancers and 251 tumor representative transcriptomes have been profiled and compared using U133A/B Affymetrix microarrays (NCBI GEO: GSE3494). The cancer samples of the patients were split into four groups (G1, G2a, G2b, G3): G1 and G3 groups with histologic grades I and II tumors, respectively; G2a and G2b groups are the sub-types of histologic grade II tumors, which have been identified based on genetic re-classification of the grade II breast cancer tissues resulting in computational pattern recognition of small and robust prognostic gene signatures (Ivshina *et al*., 2005). The order of G1, G2a, G2b and G3 corresponds to aggressiveness of breast cancer.

For every probesets presenting cis-antisense gene pair we calculated rank Kendall τ correlation coefficient of expression levels measured in the group of the breast cancer patients. We calculated correlation coefficients for all cis-antisense gene pairs and estimated statistical significance of these coefficients for each of 4 groups. Then we calculated the number of the positive and negative correlation coefficients for the 4 groups and the number of statistically significant coefficients (at fixed levels of $p = 0.01$

and $p = 0.05$). To simulate background value of correlation coefficients the same procedure was repeated for random gene pairs in the human genome.

RESULTS AND DISCUSSION

The total number of significant within-pair correlation coefficients in the expression levels of our cis-antisense paired genes was much larger than the number of the correlation expected by chance. We studied separately the subset of well-mapped sense-antisense pairs. This subset represents 182 SAT pairs which have at least one locus common for exons of the given gene pair and all the overlapped exons were mapped by at least one Affymetrix sequence. This small sub-set of cis-antisense pairs we selected in order to maximize a reliability of detection of complex signals from splice variant isoforms of a gene. Table 1 presents the numbers of sense-antisense pairs which exhibit correlations of gene expression values common for all 4 groups of patients (G1, G2a, G2b, G3). Table 1 shows numbers of positive ($\tau > 0$) and negative ($\tau < 0$) correlations within sets. It also shows that only positive correlated pairs are significant ($p < 0.005$). Moreover, the number of significant ($p < 0.05$) negative correlations does not differ from the number expected by chance (data not shown).

Table 1. Number of correlation coefficients for SAT pairs found in 4 groups

	$\tau > 0$	$\tau < 0$	$\tau > 0$ ($p < 0.005$)	$\tau < 0$ ($p < 0.005$)
Subset of SAT pairs	71	5	12	0
All SAT pairs	1758	284	352	0

Thus, the number of significant positive co-regulated pairs dominates over negatively co-regulated pairs. Therefore, co-regulation of paired genes is the most prominent type of cis-antisense pair expression pattern in different breast cancer cell types. Our results agree with recent observations in mouse transcriptome which demonstrate frequent concordant regulation of sense/antisense pairs (Katayama, 2005).

We furthermore identified two distinct sets of positively-correlated cis-antisense overlapped transcripts: permanently co-regulated gene pairs which exhibit positive correlation across G1, G2a, G2b and G3, and pairs specifically associated with each of the four groups. For instance, the RAF1/MKRN2 and CKAP1/POLR2I gene pairs represent the first set; interestingly, RAF1 is a key oncogene while MKRN2 is a transcription factor. The CR590216/EAP30 pair represents the second set (significant only in G3 group; $p < 0.05$). EAP30 can be involved in the de-repression of transcription by RNA polymerase II (Schmidt *et al*., 1999).

Our analysis reveals extraordinary reproducible positively co-regulated patterns for almost all cis-antisense loci. This allows us to suggest that cis-antisense transcripts might be controlled by global mechanisms associated with chromatin remodeling and/or mutual de-repression of synthesis of ribonucleic acids due to temporal RNA initiation of the triplex-forming DNA transcription bubble. Fig. 2 shows a hypothetical scheme, which illustrates the second model.

Based on this model we assume that gene expression is a pulse random process and that transcription of mRNA from a given strand could facilitate initiation of transcription of a gene on the opposite strand. A positive correlation (co-regulation) of expression of the gene pairs on opposite strands might be explained by direct interaction of short RNA forming temporal helical structure containing three strands (Frenster, 1965). We assume that triplex forming poly(A)-negative RNA segments of SAT could establish a locally denaturized “bubble” (Fig. 2). As polymerase (PolII) advances along the opposite strand, the RNA segment is displaced/ destroyed by RNAase(s) and then two DNA strands can be re-annealed. This model predicts a periodic re-expression of transcripts of the both sense-antisense genes.

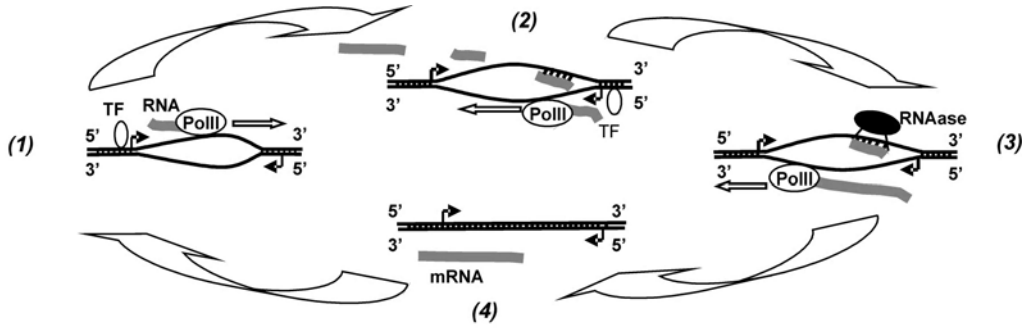


Figure 2. A hypothetical model of transcription activation of antisense mRNA by sense RNA segment.

REFERENCES

Chau Y.M., Pando S., Taylor H.S. (2002) HOXA11 silencing and endogenous HOXA1 antisense ribonucleic acid in the uterine endometrium. *J. Clin. Endocrinol. Metab.*, **87,** 2674–2680.

Ivshina A.V. *et al*. (2005) Stem cells, senescence and cancer. Keystone symposia. Abstract book, p.76.

Frenster J.H. (1965) A model of specific de-repression within interphase chromatin. *Nature*, **206**, 1269–1270.

Katayama S., Tomaru Y., Kasukawa T. *et al*. (FANTOM Consortium) (2005) Antisense transcription in the mammalian transcriptome. *Science*, **309**(5740), 1564–1566.

Schmidt A.E., Miller T., Schmidt S.L., Shiekhattar R., Shilatifard A. (1999) Cloning and characterization of the EAP30 subunit of the ELL complex that confers derepression of transcription by RNA polymerase II. *J. Biol Chem.*, **274**, 21981–21985.

Vanhee-Brossollet C., Vaquero C. (1998) Do natural antisense transcripts make sense in eukaryotes? *Gene*, **211**, 1–9.

THE SITEGA AND PWM METHODS APPLICATION FOR TRANSCRIPTION FACTOR BINDING SITES RECOGNITION IN EPD PROMOTERS

***Levitsky V.G.*[*1, 2], *Ignatieva E.V.*[1, 2], *Ananko E.A.*[1], *Merkulova T.I.*[1]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] corresponding author: e-mail: levitsky@bionet.nsc.ru

Key words: transcription factor binding sites recognition, large-scale genome research, position weight matrix, discriminant analysis, genetic algorithm

SUMMARY

Motivation: Development of methods to predict functional transcription factor binding sites (TFBSs) is very important for eukaryotic genes annotation. But the high false positive rate is a serious issue in attempts to reliably predict TFBSs.

Results: We propose the combined approach to search for TFBSs. The approach compiled the SiteGA method that takes into account the interactions between different TFBS positions and position weight matrix (PWM) method. Both methods have been applied to four transcription factor (TF) types (IRF1, PPAR, SREBP and SF-1). The combined approach was tested on the set of the promoters from EPD database. The approach application allows revealing most reliable potential TFBS targets.

Availability: http://wwwmgs2.bionet.nsc.ru/mgs/programs/sitega/.

INTRODUCTION

Recognition of TFBSs by computer methods is an effective approach to the search and analysis of the regulatory gene regions. Widely used PWM model for DNA binding implies that there is some contribution from each base at each position and that the sum of all the contributions is above a certain threshold (Stormo, 2000). Nevertheless the weight matrix is severely limited by the assumption that positions in a binding site (BS) contribute additively to the total score (Benos *et al.*, 2002; Zhou, Liu, 2004). As a result, the accuracy of the recognition sometimes is far too low for large-scale genome research. To overcome this drawback we combined the PWM method with SiteGA, which takes into account the interactions between different TFBS positions. We revealed that combined recognition might significantly potentate the recognition power.

METHODS AND ALGORITHMS

We used SiteGA (Levitsky *et al*., 2006) and PWM (Stormo, 2000) site recognition methods in our analysis. Samples of nucleotide sequences used in analysis are presented in Table 1. The train data sets (sequences with flanks with centrally located BS) were extracted from the TRRD database (Kolchanov *et al.*, 2002). The SiteGA and PWM methods used the train sequences of almost the same length. The control BS sets (IRF1, PPAR, SREBP) were derived from TRRD and literature sources and were used for

thresholds setting. For both methods these settings corresponded to 50 % of true positive rates, which were estimated by the control BSs sets. To set thresholds for SF-1 BSs, we used two samples of vertebrate promoters, extracted from TRRD: (i) genes controlling steroidogenesis and their orthologs (STER+), this sample didn't contain train BSs; (ii) the remaining genes which lack the experimentally approved SF-1 sites (STER-). The thresholds settings for SF-1 BSs were based on the restriction of predicted sites portion for set STER+ (i.e. approximately 20 % of sequences contained potential sites). The set of human promoters extracted EPD database (Schmid *et al.*, 2006) were finally used for potential TFBS targets search. Thus for all TF types both methods (SiteGA & PWM) used coordinated by true positive rates thresholds settings. The estimates for false positive (FP) rates calculated by the random sequences (train data with preserved nucleotide content) at specified above thresholds are given in Table 1.

Table 1. Samples of nucleotide sequences

Sample type	Sample name	Sequence length, nt	No. of sequence			FP rate	
			Train	Control	Test	SiteGA	PWM
TFBSs	IRF1	56	30	29		3.0E-07	3.0E-06
	SF-1	25, 30[3]	54			2.6E-05	5.7E-05
	PPAR	25	54	16		2.9E-04	1.6E-04
	SREBP	18	38	8		6.1E-04	1.1E-03
Promo-ters	STER+, [-350;+50][1,2]	400		70			
	STER-, [-350;+50][1,2]	400		1285			
	EPD, [-550;+50][1,2]	600			1871		

[1] – location relative to transcription start site; [2] – lacking the 5′- or 3′-flanks of nucleotide sequences completed with the symbol "n"; [3] – for SiteGA and PWM methods training, correspondingly.

We applied nucleotide and dinucleotide PWM methods as follow.

$$w_{i,j} = -n_{i,j} \times \ln(\mathrm{p}_j). \quad (1)$$

Here n_{ij} is the count of nucleotide (dinucleotide) j in position i and p_j is nucleotide (dinucleotide) j frequency for train set. We chose for each BS the best PWM type (nucleotide or dinucleotide). Only for SREBP BS nucleotide matrix was used as most accurate, for other BSs we used dinucleotide matrix.

The training window length and location search implied the search among different sizes (from 10 to 60 nt) and slightly shifted locations (Fig. 1), i.e. for each window size three locations were tested. The recognition accuracy estimate based on the standard jack-knife test was used for window size and location selection.

Figure 1. The scheme used for PWM method's training window length and location search.

The SiteGA method was implemented using of a genetic algorithm (GA) involving a discriminant analysis. The GA handled a population of individuals, which were defined as sets of N locally positioned dinucleotides (LPDs). Each LPD was specified by it location (a, b) within the analyzed window and type d_j (j = 1, …, 16). The initial GA population

consisted of individuals of arbitrarily assigned LPDs. Then GA produced iterative mutations (Fig. 2*a*, *b*) and recombinations (Fig. 2*c*). The GA was based on the fitness maximization. Let us consider the real (1) and random (2) (obtained by shuffling of the real sequences) sequences sets. The fitness of an individual was given by the Mahalanobis distance $R^2(\{f_n\})$.

$$R^2(\{f_n\}) = \sum_{k=1}^{N}\sum_{n=1}^{N}\{[f_n^{(2)} - f_n^{(1)}] \times S_{n,k}^{-1} \times [f_k^{(2)} - f_k^{(1)}]\}. \quad (2)$$

Here, N is total No. of LPDs, $f_n^{(1)}$ and $f_n^{(2)}$ are mean frequencies of the nth LPD calculated for the real and random sets respectively; $S_{n,k}^{-1}$ is an element of the matrix $|S^{-1}|$ inverse to the matrix $|S|=|S^{(1)}|+|S^{(2)}|$. These are the covariance matrices of the vectors of LPDs over the sequence sets 1 and 2.

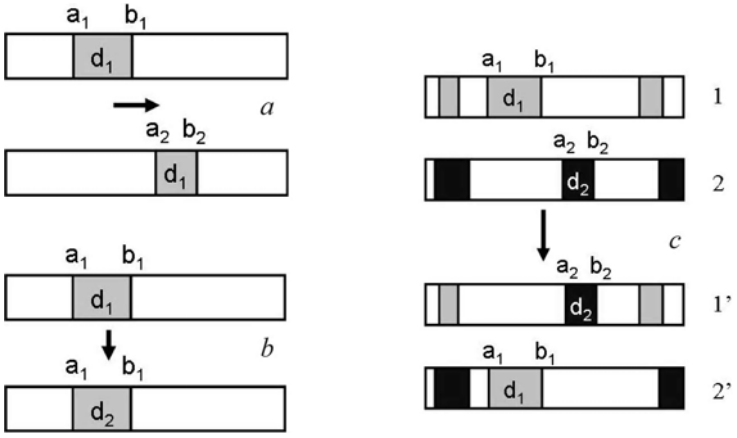


Figure 2. The elementary GA operations: *a*, *b* – mutations, *c* – recombination. *a* – change of LPD location ($[a_1, b_1] \rightarrow [a_2, b_2]$, the dinucleotide type d_1 remains the same); *b* – change of LPD dinucleotide type ($d_1 \rightarrow d_2$, the location $[a_1, b_1]$ remains the same). *c* – exchange of two LPD $\{[a_1, b_1], d_1\}$ and $\{[a_2, b_2], d_2\}$ between two parent individuals (1 & 2), 1' and 2' – daughter individuals; LPD belonging to parent individuals corresponds to light and dark colors.

The recognition function value was calculated for a nucleotide sequence X as follow.

$$\varphi(X) = \frac{1}{R^2} \times \sum_{n=1}^{N}\sum_{k=1}^{N}\{[f_n(X) - (1/2) \times [f_n^{(2)} + f_n^{(1)}]] \times S_{n,k}^{-1} \times [f_k^{(2)} - f_k^{(1)}]\}. \quad (3)$$

The combined SiteGA & PWM approach application implies the obligatory BS recognition by both methods.

RESULTS AND DISCUSSION

The results of test EPD data analysis are given in Table 2.

For SREBP BS both methods predicted nearly equal portions of sites (567 and 583). For IRF1 BS SiteGA was able to yield very small portion of predicted sites in comparison with PWM (29 against 95). For other BS (SF-1 and PPAR) differences in

predicted sites portions were not very noticeable. Our approach appeared to be the most accurate for IRF1 and SF-1 BSs (Table 1): in every case totally in whole EPD promoter set (1871 sequences) less than 30 putative sites were found (Table 2). Totally 25 (1.5 %) and 20 (1 %) of EPD promoters contained correspondingly IRF1 and SF-1 predicted sites. That observation may be very promising for large-scale genome analysis, since human genome contains at least tenfold greater number of genes than analyzed here EPD set.

Table 2. The analysis of predicted TFBS density in EPD promoters by SiteGA, PWM, and combined SiteGA & PWM methods

TF type	SiteGA	PWM	SiteGA & PWM
IRF1	29 (3.2E-05)	95 (1.0E-04)	27 (2.6E-05)
SF-1	75 (7.7E-05)	148 (1.5E-04)	20 (2.1E-05)
PPAR	315 (3.7E-04)	192 (2.0E-04)	106 (1.1E-04)
SREBP	567 (5.8E-04)	583 (5.9E-04)	110 (1.1E-04)

For each TF type the No. of predicted sites and its ratio to the total count of analyzed positions are given.

The promoters which containing predicted by combined approach IRF1 and SF-1 BSs are given in Table 3. It may be concluded that combined method application allows significantly decrease false positive rate (Table 2) and it application gives opportunities to reveal the most reliable potential TFBS targets (Table 3).

Table 3. Human promoters annotated in EPD which containing predicted IRF1 and SF-1 BSs by combined SiteGA & PWM approach

Functional class	SF-1	IRF1
Known target	CG/LH/FSH/TSH-α, CGI127, HSD3B2	2'5'-oligoAsynt., complement f. B, IFNα 13, β-interferon, HLA B, IFI27, CEACAM1, SP100, IFI 54K, IFI 6-16
Very possible target	PBGD E E2P2, CDC42EP2, TRAP1, ACPP, PTGES2	BST2, APOL1
Possible target	FXR1, FXYD3, VAPA	IL-4 (BSF-1), IL-5 (EDF/TRF), APOL3, TAPBP, POLD2, EIF3S7, PRG1, PHGDH, GLRX, C1QBP
LRH-1 related (SF-1 only) [1]	Glucagons, CTRB1, CPA2	
Unknown	TCR vα HD-Mar, ATP6V0D1, RPS5, HNRPH2, SAT, STMN2	Link, SDHD, RPL35A P1

[1] – LRH-1 is a close homolog of SF-1 (Fayard *et al.*, 2004).

ACKNOWLEDGEMENTS

The work was supported by the RFBR (grants Nos 05-04-49111, 05-07-98012); U.S. Civilian Research and Development Foundation for the Independent States of the Former Soviet Union (CRDF) and the Ministry of Education of Russian Federation within the Basic Research and Higher Education Program (Award No. REC-008, grant Y2-B-08-02). The authors are grateful to Dr. Lokhova I.V. for technical support, and to Dr. Podkolodny N.L. for helpful discussions.

REFERENCES

Benos P.V. *et al.* (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucl. Acids Res.*, **30**, 4442–4451.

Fayard E. *et al.* (2004) LRH-1: an orphan nuclear receptor involved in development, metabolism and steroidogenesis. *Trends in Cell Biol.*, **14**, 250–260.

Kolchanov N.A. *et al.* (2002) Transcription Regulatory Regions Database, (TRRD): its status in 2002. *Nucl. Acids Res.*, **30**, 312–317.

Levitsky V.G. *et al.* (2006) Method SiteGA for transcription factor binding sites recognition. *Biofizika* (in press).

Schmid C.D. *et al.* (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucl. Acids Res.*, **34**, 82–85.

Stormo G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Zhou Q., Liu J.S. (2004) Modelling within-motif dependence for transcription factor binding site predictions. *Bioinformatics*, **20**, 909–916.

DETECTING HAIRPINS IN 3′-UNTRANSLATED REGIONS OF HIGHLY EXPRESSED GENES IN ACTINOBACTERIA

Lyubetskaya E.V., Seliverstov A.V.*, Lyubetsky V.A.
Institute for Information Transmission Problems, RAS, Moscow, Russia,
*Corresponding author: e-mail: slvstv@iitp.ru

Key words: transcription termination, bacteria, DNA relaxation, highly expressed genes

SUMMARY

Motivation: In bacterial genomes antiparallel genes are common within one chromosome. At least in cases when one of the genes is highly expressed, their shared 3′-untranslated region should be involved in transcription termination. At the same time, gene transcription very often entails the formation of DNA supercoils at this site.

Results: Very long hairpins are found in some Actinobacteria downstream of highly expressed genes. But these hairpins do not resemble known types of terminators involved in expression regulation. We suppose that they are involved in DNA relaxation and an uncharacterized termination mechanism.

INTRODUCTION

In Actinobacteria, gene expression is typically regulated on translation level through overlapping of ribosome binding site (Seliverstov *et al*., 2005), whereas in gamma- and alpha-proteobacteria the regulation follows classical attenuation scenario (Vitreschak *et al*., 2004). In most well studied bacteria, like *Escherichia coli* and *Bacillus subtilis*, operons end with transcription terminators, GC-rich hairpins with adjacent poly-U tract downstream. However, such terminators are rare or absent altogether from some bacterial taxa, like e.g. Cyanobacteria and *Mycobacterium* (Washio *et al*., 1998; Unniraman *et al*., 2002). One might suppose that transcription termination in Actinobacteria involves alternative secondary structures of double-stranded DNA.

In this work we sought for DNA structures putatively responsible for termination in 3′-untranslated regions of highly expressed genes encoding tRNA, elongation factors and some important proteins.

Structures found in 3′-untranslated regions of these genes allowed for better defining operon boundaries and predicting highly transcribed DNA regions (this problematic was discussed in detail in (Ishchukov *et al*., 2004)). Particularly, results of the search algorithm were used to reveal expressed paralogs.

RESULTS

Bacterial genomes were obtained from GenBank. Long hairpins are found in 3′-untranslated regions of genes encoding tRNA and some proteins, being especially abundant in intergenic spaces between antiparallel genes, with one of them coding for tRNA. For example, in *Propionibacterium acnes* long hairpins are found downstream of

tRNA-Ala (*ppa2421*), tRNA-Arg (*ppa2413*), tRNA-Arg (*ppa2189*), tRNA-Asn (*ppa2422*), tRNA-Glu (*ppa2432*), tRNA-Lys (*ppa0181*), tRNA-Lys (*ppa1961*), tRNA-Met (*ppa2423*), tRNA-Phe (*ppa2454*), tRNA-Pro (*ppa2428*), tRNA-Thr (*ppa2412*). In *Corynebacterium efficiens* long hairpins are found downstream of tRNA-Ala, tRNA-Arg, tRNA-Asp, tRNA-Leu, tRNA-Pro, tRNA-Ser. Moreover, such hairpins are found downstream of other six highly expressed protein-coding genes in *P. acnes*, six such genes in *C. efficiens* and five protein-coding genes in *Mycobacterium bovis*.

A part of our data is presented in Table 1:

Table 1. The numbers of hairpins with length equal or higher than L for leader regions (1), regions of converging located genes (2), coding regions (3), regions of divergently located genes (4) are shown in second, third, forth and fifth columns, respectively

L	(1)	(2)	(3)	(4)	L	(1)	(2)	(3)	(4)
Corynebacterium efficiens					Mycobacterium bovis				
25	2	1	0	0	25	2	12	1	1
23	2	2	0	0	23	3	15	1	1
20	6	16	0	0	20	4	17	2	1
17	23	37	5	1	17	12	21	6	2
15	44	57	13	6	15	17	22	28	3
10	182	121	960	27	10	188	37	1499	37
Corynebacterium glutamicum					Mycobacterium leprae				
25	0	1	0	0	25	5	0	2	1
23	0	3	1	0	23	5	0	2	1
20	2	16	1	0	20	8	1	4	1
17	12	40	4	0	17	10	1	12	1
15	29	59	8	0	15	14	1	27	2
10	221	133	617	29	10	111	4	482	8
Corynebacterium diphtheriae					Mycobacterium avium				
25	0	2	0	0	25	1	1	0	0
23	1	6	0	0	23	3	1	0	0
20	2	14	0	0	20	4	4	0	0
17	14	32	2	0	17	8	8	8	0
15	24	48	7	0	15	17	19	24	3
10	137	76	497	23	10	257	46	2197	66
Propionibacterium acnes					Mycobacterium tuberculosis				
25	0	0	0	0	25	4	11	3	1
23	0	0	0	0	23	5	14	3	1
20	0	2	1	0	20	7	17	3	2
17	3	17	2	0	17	17	20	6	3
15	8	27	3	0	15	24	21	24	5
10	101	63	571	17	10	202	36	1519	30

The length of a hairpin is the number of nucleotides in its shoulders. The numbers of hairpins with length equal or higher than L for leader regions (1), trailer regions (2), coding regions (3), regions of divergently located genes (4) are shown in second, third, forth and fifth columns, respectively. Besides, the Table 1 shows only hairpins with loops shorter than 15 nucleotides and with only one internal loop 2 nucleotides or less in length. Moreover, the left shoulder was not allowed to contain regions complementary to those of the hairpin loop.

Hairpins of 18–27 bp length (called abnormally long hairpins) are seldom found in some genomes (results shown for *P. acnes*). For each gene, its leader region was defined as a region no more than 300 bases in length and not crossing the bounds of neighbor genes. Transcription initiation site was not considered and is usually unknown. In the *P. acnes* genome, mass searches for long hairpins without bulges in 5′-untranslated regions of up to 300 bp length upstream of all genes contained in GenBank annotation resulted in detecting four hairpins with stem size exceeding 18 bp. Hairpins with stems longer than 28 bp were not detected in

intergenic spaces of this genome. Two hairpins were found with 27 bp, one – with 22 bp and one – with 18 bp-long stems. Here the first two are described.

A hairpin with a 4 bases-long loop and 27 bp-long stem without bulges was detected in 3′-untranslated region immediately following the stop codon of elongation factor G at an 8-base distance. Downstream of the hairpin the gene of transmembrane protein PPA1874 is located. Both genes are of considerable length.

The other hairpin is confined in between genes *ppa1754* and *ppa1753* encoding the alpha subunit of highly expressed succinyl-CoA synthetase and a putative transmembrate protein, respectively.

DISCUSSION

In bacterial genomes antiparallel genes are common within one chromosome. At least in cases when one of the genes is highly expressed, their shared 3′-untranslated region should be involved in transcription termination, which is probably mediated by the found hairpins. For instance, tRNAs genes are highly expressed because of intensive usage of their products in the cell.

Besides, gene transcription entails formation of DNA supercoils, also in the 3′-untranslated region, which are conventionally thought to be relaxed by topoisomerases. Although, in intergenic regions, with at least one gene highly expressed, an alternative process might be involved in DNA relaxation with the use of detected hairpins.

In other words, in 3′-untranslated region of a highly expressed gene (especially if it belongs to a pair of antiparallel genes) one might expect to find a pair of hairpins forming the so called “cross” on two DNA strands.

Comparative analysis of hairpins in orthologs of close species reveals high divergence of their primary structure and high conservation of topology, which implies severe functional constraints imposed on the hairpin secondary structure.

These hairpins do not resemble known types of terminators involved in expression regulation. Indeed, the Rho-independent terminator typically contains a U-rich region, and the Rho-protein binding site has a UC-rich region lacking hairpins. None of these is found in or nearby the detected hairpins.

ACKNOWLEDGEMENTS

The authors are grateful to L.Y. Rusin and A.G. Vitreschak for discussion and help. This study was supported by ISTC grant No. 2766.

REFERENCES

Ishchukov I.M. *et al*. (2004) A new algorithm for recognizing the operon structure of procaryotes. *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure*, Novosibirsk, pp. 73–76.

Seliverstov A.V. *et al*. (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiol*., **5,** 54.

Unniraman S. *et al*. (2002) Conserved economics of transcription termination in eubacteria. *Nucl. Acids Res.*, **30**, 675–684.

Vitreschak A.G. *et al*. (2004) Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. *FEMS Microbiol Lett*., **234**, 357–70.

Washio T. *et al*. (1998) Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. *Nucl. Acids Res.*, **26**, 5456–5463.

MODELING CLASSIC ATTENUATION REGULATION OF GENE EXPRESSION IN BACTERIA

Lyubetsky V.A., Seliverstov A.V.*
Institute for Information Transmission Problems, RAS, Moscow, Russia
* Corresponding author: e-mail: slvstv@iitp.ru

Key words: classic attenuation regulation, transcription elongation, RNA-polymerase termination, translation initiation and elongation, energy of secondary structure, transition between secondary structures

SUMMARY

Motivation: Attenuation regulation, particularly, in its classic form is well described on comparative genomic level using evidence from both datamining and experiment. Even being confronted with difficulties in choosing adequate parameter settings, developing a rigor and effective computer model of each attenuation type is a timely and important task. Such a model is prerequisite to *in-silico* choose between alternative hypotheses of any gene leader region, as well as to study attenuation mechanisms in conjunction and in comparison with the mechanism of protein-DNA interaction (repressor-activator).

Results: An effective computer model of classic attenuation regulation is developed. The model is based on rigor and explicit statements (viz., description of all correlations and parameter value settings), which provides for its greater accuracy to explain experimental data. Results of computations reveal qualitatively correct correlations between termination probability and amino acid concentration for leader regions with predicted attenuation. When applied to random sequences, the model produces correlation values varying around a certain constant, which indicates the lack of regulation.

MODEL

The approach is based on modeling RNA secondary structure in the regulatory region between the ribosome and RNA-polymerase, resonant equations of the RNA-polymerase inhibition by helices in this region (equation for F given bellow), modeling transcription and translation initiation and elongation. Microstate is a set of continuous fragments, referred to as *hypohelices*, of any non-continuable *helices* in the same region. The model describes transitions between microstates: decomposition and binding of *hypohelices* in the same region.

The rate constant of transition between microstates ω and ω' is calculated as follows:

$$K(\omega \to \omega') = \kappa \cdot \exp[\frac{1}{2}\Big((G_{loop}(\omega) + G_{hel}(\omega)) - (G_{loop}(\omega') + G_{hel}(\omega'))\Big)] .$$

Here $R \cdot T \cdot (G_{loop}(\omega) + G_{hel}(\omega))$ is free energy of RNA secondary structure (loops and helices) in microstate ω, R – *universal* gas constant, T – *absolute* temperature (for details ref. to Mironov, Lebedev, 1993).

Two microstates ω and ω' belong to one *macrostate* Ω if both ω and ω' are realized by identical diagram (for definitions of *diagram* and its *chord* ref. to Lyubetsky *et al*., 2006); intuitively it means that both $K(\omega \to \omega')$ and $K(\omega' \to \omega)$ are relatively large. It was our aim to achieve that transitions between two microstates ω and ω' within any macrostate Ω are fast, and those between any microstates ω and ω' from different Ω and Ω', respectively, are slow.

Absolute probabilities of transitions between microstates ω and ω' in macrostate Ω are inessential in our model. Instead, transitions in the set of all microstates ω in any macrostate Ω are required to produce Boltzmann-Gibbs stationary probability distribution:

$$p(\omega) = \frac{\exp\left(-(G_{loop}(\omega) + G_{hel}(\omega))\right)}{z(\Omega)}, \text{ where}$$

$$z(\Omega) = \sum_{\omega \in \Omega} \exp\left(-G_{loop}(\omega) - G_{hel}(\omega)\right).$$

Trivial averaging over all pairs of microstates ω in Ω and ω' in Ω' produces the following equation for the transition rate constant between macrostates Ω and Ω' that applies to both increase and decrease of macrostate by *one chord*: $K(\Omega \to \Omega') = \sum_{\omega \in \Omega} \sum_{\omega' \in \Omega'} p(\omega) \cdot K(\omega \to \omega')$. All other transitions between macrostates are null.

The rate constant of polymerase transition from one nucleotide to the next is calculated as $\nu(\Omega) = \overline{\lambda}_{pol} - F(\Omega)$, where Ω is a macrostate, and $F(\Omega)$ is effective decrease of the polymerase rate constant in s^{-1}. In the model, polymerase deceleration by hairpin ω is described as follows:

$F(\omega) = \frac{\delta}{L_1^2 \cdot (p(\omega) - p_0)^2 + 1} \cdot \exp\left(-\frac{r}{r_0}\right)$, where *r* is distance between the terminus of hairpin and the polymerase. *Parameters* L_1, p_0, r_0, δ depend on polymerase characteristics and value $p(\omega)$ – on hairpin ω. For a hairpin consisting of the handle and the loop, *p* is estimated from the equation: $\mathrm{tg}(p \cdot h) = \frac{2}{p \cdot l}$, $0 < p \cdot h < \frac{\pi}{2}$, where *h* is *handle length*, i.e. the number of its base pairs, and *l* is *loop length*. An analogous equation is used for an arbitrary hairpin.

The rate constant of the polymerase sliding within a T-rich region is estimated as $\mu(\Omega) = F(\Omega)/4$ (Yin *et al*., 1999).

On non-regulatory codons, the rate constant λ_{rib} of ribosome elongation by 1 nucleotide is $\overline{\lambda}_{rib} = 45 s^{-1}$. On regulatory codons, λ_{rib} depends on concentration *c* of aminoacyl-tRNA according to the Michaelis-Menten law: $\lambda_{rib}(c) = \frac{\overline{\lambda}_{rib} \cdot c}{c_0 + c}$.

To model obstacles in ribosome binding, we incorporated ribosome binding rate constant $K_0 = \lambda_0 \cdot \frac{d_{open}}{d_{\max}}$, where d_{open} is current value of the maximum number of open nucleotides in the Shine-Dalgarno sequence (provided that the start codon is open), $d_{\max}$ – the length of the sequence and λ_0 – translation initiation parameter.

Standard Monte-Carlo technique is used in modeling. For example, neighborhood of given state Ω, centered in Ω, is a set of all states Ω' with non-zero probability of transition from Ω to Ω' by both increase and decrease of macrostate Ω. If given neighborhood contains *n* states and corresponding transition rate constants are k_1, ..., k_n, the next state on the trajectory of transitions is determined by realizing random variable $i \to \frac{k_i}{\sum k_i}$.

The following parameter settings were chosen: $\kappa = 10^3$, r_0 within the range 2–8, $L_1 = 14.5$, $p_0 = 0.167$, $\delta = 25$, $\kappa = 10^3$ s^{-1}, $c_0 = 1$. "Sizes" of ribosome and polymerase are $s_0 = 12, s_1 = 5$.

The purpose of modeling was estimating function $p = p(c)$ of correlation between termination probability and concentration c of amino acid or concentration c of aminoacyl-tRNA synthetase for operon leader regions in bacteria. These estimated were also obtained for random sequences (see below). Function $p(c)$ was estimated with repeating the modeled process certain number of times (usually 10^3–10^4) under given c increment, and $p(c)$ was calculated as a fraction of times when termination occurred.

Computer assays were "positive" when all available regions with putative attenuation (using evidence e.g. from Vitreschak *et al*., 2004) were analyzed under fixed values of the above described parameters. The assays were "negative" under the same parameter settings when modeling was done with "*random*" *sequences* assembled from the leader peptide gene upstream of *trpE* in *Vibrio cholerae*, a U-rich terminator from the same leader region and a random sequence in four-letter alphabet of random length intercalating the two. Positive assays were expected to return approximately monotonous growth of function $p(c)$, while negative – to demonstrate its absence. All positive assays, except for tryptophan biosynthesis operons in *Streptomyces* spp., returned approximately monotonous growth (ref. to Results), and all negative – oscillations around different constants.

RESULTS AND DISCUSSION

Values in the Table were obtained by computing with our model on leader regions upstream of gene *trpE* in *Corynebacterium diphtheriae*, *Corynebacterium glutamicum*, *Agrobacterium tumefaciens*, *Bradyrhizobium japonicum*, *Rhodopseudomonas palustris*, *Rhizobium leguminosarum*, *Sinorhizobium meliloti*, *Escherichia coli*, *Vibrio cholerae*, and also *for* gene *trpS* in *Streptomyces avermitilis*. The results are in congruence with multiple alignments of corresponding leader regions, which are available in publications for actinobacteria (Seliverstov *et al*., 2005) and proteobacteria (Vitreschak *et al*., 2004).

For *C. glutamicum*, termination probability estimated in the model doubles under tryptophan concentration growth but still was very low. For some alpha-proteobacteria, modeled termination probability increases considerably: 48-fold in *R. palustris*, 7.6-fold in *S. meliloti*, and 16.6-fold in *V. cholerae*. The ranges decrease under κ growth, and in this sense their interpretation is unclear.

Some rows of the Table represent not strictly monotonous pattern. This might be accounted for by precision of modeling being below 0.01–0.02, which also depends on characteristics of the random seed generator. Classic attenuation is applicable within specific intervals of the c value that are determined individually for each gene and organism. Small size of such interval measured in the model in c / c_0 does not necessarily imply small physical values, e.g. in mM/l. Also, bacteria in favorable natural environment do not display strictly monotonous function $p(c)$. The results (presented partially) reveal

correlation $p(c)$ congruent on the qualitative level with presence of attenuation in most gene leader regions studied. All negative assays (data not shown) returned $p(c)$ values oscillating around certain constant.

Thus, our model can be used to predict the impact of point mutations in regulatory regions on attenuation regulation and to predict stability of this system during the course of evolution. It can also be incorporated into a broader non-linear model of bacterial metabolism with dynamic modeling of gene expression regulation. Another possible application of the model is prediction of attenuation regulation by modeling correlation between the enzyme activity and amino acid concentration for a *single sequence*, thus eliminating the need to analyze sequence profiles.

A method is proposed to objectively choose model settings on the basis of source data. The computer program offers high flexibility to vary all model parameters and correlations. The model was applied to biological data to assess its relative robustness against varying parameter settings and to obtain their estimates using Monte-Carlo approximations of typical stem lengths, macro- and microstate ratios, lengths of the ribosome and polymerase neighboring-state transition cycles, etc.

Table 1. Termination probability $p(c)$ against concentration c of triptophanyl-tRNA in various bacteria

Species	Concentration c										
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**0.25**	**0.30**	**0.35**	**0.40**	**0.45**	**0.50**
C. diphtheriae	0.34	0.34	0.39	0.46	0.50	0.54	0.53	0.53	0.53	0.52	0.54
C. glutamicum	0.05	0.06	0.08	0.10	0.10	0.09	0.09	0.09	0.10	0.10	0.10
S. avermitilis, trpS	0.06	0.13	0.21	0.26	0.28	0.29	0.30	0.30	0.32	0.32	0.30
A. tumefaciens	0.49	0.50	0.62	0.70	0.74	0.78	0.77	0.78	0.82	0.80	0.79
B. japonicum	0.19	0.20	0.24	0.26	0.28	0.26	0.26	0.27	0.26	0.26	0.26
R. leguminosarum	0.23	0.30	0.42	0.55	0.60	0.65	0.67	0.70	0.71	0.71	0.71
R. palustris	0.01	0.22	0.40	0.48	0.56	0.59	0.60	0.60	0.63	0.61	0.62
S. meliloti	0.07	0.11	0.23	0.37	0.43	0.49	0.48	0.51	0.50	0.53	0.51
E. coli	0.34	0.46	0.54	0.68	0.70	0.70	0.71	0.73	0.75	0.75	0.74
V. cholerae	0.05	0.16	0.39	0.57	0.70	0.74	0.77	0.77	0.80	0.79	0.81

ACKNOWLEDGEMENTS

The authors are grateful to A.A. Mironov and L.Y. Rusin for discussion and help. This work was partly supported by ISTC 2766.

REFERENCES

Lyubetsky V.A., Rubanov L.I., Seliverstov A.V., Pirogov S.A. (2006) Model of genes expression regulation in bacteria by means of formation of secondary RNA structures. *Mol. Biology*, **40**, 497–511.

Mironov A.A., Lebedev V.F. (1993) A kinetic model of RNA folding. *BioSystems*, **30**, 49–56.

Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology*, **5**, 54.

Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. (2004) Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis. *FEMS Microbiology Letters*, **234**, 357–370.

Yin H., Artsimovitch I., Landick R., Gelles J. (1999) Nonequilibrium mechanism of translation termination from observations of single RNA polymerase molecules. *PNAS*, **96**, 13124 – 13129.

STRUCTURAL VARIANTS OF BINDING SITES FOR GLUCOCORTICOID RECEPTOR AND THE MECHANISMS OF GLUCOCORTICOID REGULATION: ANALYSIS OF GR-TRRD DATABASE

Merkulov V.M.**, Merkulova T.I.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: merkti@niboch.nsc.ru

Key words: glucocorticoid receptor, binding sites, gene regulation, database

SUMMARY

Motivation: Glucocorticoid receptor (GR) is an important regulator of many genes involved in a variety of biochemical and physiological processes. The features of the structure and localization of DNA binding sites for GR may be significant for ensuring specificity of glucocorticoid-mediated regulation of different genes.

Results: GR-TRRD database accumulates the largest out of currently published samples of nucleotide sequences that are experimentally proved to bind GR (glucocorticiod receptor binding sites, GRbss). This sample consists of 152 GRbss from 77 genes controlled by glucocorticoids. Analysis of the sample has shown that the structure of only half of GRbss (53 %) corresponds to traditional viewpoint about structural organization of glucocorticoid response element (GRE) as an inverted repeat of hexameric half-site sequence TGTTCT (Aranda, Pascual, 2001). 40 % of GRbss contain only hexameric half-site. Notably, there exist experimental evidence about participation of most of these GRbss in glucocorticoid regulation. As a result of increasing the number of sequences in the sample of GRbss, we have specified the consensus of sites organized in a form of inverted repeat (palindromic GREs). Also, possible mechanisms of action of hexameric half-sites in glucocorticoid induction have been discussed.

Availability: (http://wwwmgs.bionet.nsc.ru/mgs/papers /merkulova/gluc/).

INTRODUCTION

Glucocorticoid hormones regulate basic vital functions of the organism in vertebrates: coordinated growth, differentiation, reproduction, adaptation, and behavior. As a rule, glucocorticoid effect in the target cells is produced by binding to a specific intracellular receptor (glucocorticoid receptor, GR) that regulates genes via direct interaction with specific DNA sequence and/or via protein/protein interactions with the other transcription factors (Schoneveld *et al*., 2004). GR is a member of nuclear hormone receptor superfamily. The basis of transcription factor binding sites of this superfamily is produced by two hexameric motifs: 1) TGTTCT (GR, mineralocorticoid, androgen, and progesterone receptors) and 2) TGACCT (the other receptors). Due to current opinion, GR, like the other steroid hormone receptors, interacts with DNA in a form of homodimer recognizing the inverted repeat TGTTCT (or TGACCT, in case of the estrogen receptor) separated by three base pairs. Thyroid hormone receptors, vitamin D receptors, retinoic acid receptors, and numerous orphan receptors (HNF4, COUP, PPAR, CAR, PXR, LXR, etc.) are united in a group of proteins, which in a form of homo- or

heterodimers interact with direct, inverted, or everted repeats of TGACCT motif with the spacer varying from 0 to 9 bp. Several orphan nuclear receptors bind DNA as monomers (SF1, LRH1, ROR, and ERR). The single TGACCT motif being the basic element for the binding of these receptors is preceded by 5′- flanking AT-rich sequence consisting of three-six nucleotide bases (Aranda, Pascual, 2001).

However, the primary structure of many glucocorticoid receptor binding sites (GRbss) from different genes is beyond the frames of the standard model. In particular, there exist GRbss participating in glucocorticoid regulation and containing only a single copy of TGTTCT hexanucleotide, to which GR binds as a monomer. In addition, the functional GRbss organized as direct hexanucleotide repeats have been detected. Also, the cases are known when GR forms a heterodimer with the other transcription factors. We aimed to elucidate the relative representation of different structural variants in the sample of 152 GRbss, experimentally found in various genes. The sample was extracted from the section GR-TRRD (Glucocorticoid-Regulated Genes TRRD) (Merkulova *et al*., 1997) of the database TRRD (Transcription Regulatory Regions Database) (Kolchanov *et al*., 2002) and contains the overwhelming majority of currently known natural GRbss.

RESULTS AND DISCUSSION

Due to common viewpoint, GRbss are organized as inverted repeats of hexanucleotide motif TGTTCT with the spacer of 3 bp, i.e., the "palindrome" AGAACAnnnTGTTCT, which interacts to GR-homodimer. These sites are considered as the classic glucocorticoid response elements (GREs), or DNA regions capable to produce glucocorticoid response (Aranda, Pascual, 2001; Schoneveld *et al*., 2004).

Based on the GR binding data, we have compiled the set of 152 nucleotide sequences of GRbss extracted from regulatory regions of 77 genes controlled by glucocorticoids. For most of them (80 %), the data are known about their functioning as GREs. 81 out of 152 sites from the sample are homologues of the "palindrome". For these sites, the number of discordant positions relative to this sequence does not exceed six, so that each of the halves of repeat has at most three discordances. Most sites from this group are characterized by very good homology with AGAACAnnnTGTTCT. For 52 of such GRbss, the number of discordances is less than three, with at most two discordances at each half-site. About 40 % of sites (62) from the sample are not the "palindromes". Their sequences contain only the hexanucleotide TGTTCT (the number of discordances varies from 0 to 2), whereas the neighboring 5′ sequence contains at most two coincident positions with the left half of the inverted repeat. Neither two sites with two coincident positions have simultaneously G nucleotide at position 2 or C nucleotide at position 4, which are crucial for binding with GR (Beato *et al*., 1989).

Besides, the sample of 152 GRbss contains 3 sites organized in a form of direct repeat of hexanucleotide TGTTCT, which binds to GR-MR heterodimer; 4 sites binding GR in the region without homology to the hexanucleotide; a single site containing two overlapping "palindromes" binding with GR-tetramer; and a single site, where "palindrome" overlaps with the hexanucleotide and binds with GR-trimer.

By increasing the number of sequences in the sample of GRbss and dividing GRbss into the structural variants, we have worked out in more details the consensus of "palindromic" sites. In Fig. 1*a*, the frequency matrix and consensus made in 1989 on the basis of analysis of 25 GRbss (Beato *et al*., 1989) are illustrated. In Fig. 1*b*, the matrix and consensus designed by us as a result of analysis 81 sites from GR-TRRD, in which both halves of the inverted repeat were found, are presented. Currently, to search for potential GRbss, two variants of "palindromic" sequence are used. The Beato consensus (Fig. 1*a*) is used more frequently than the perfect inverted repeat AGAACAnnnTGTTCT.

As follows from our results, by increasing the number of sequences in the sample, the consensus of GRbss approximates to the perfect inverted repeat.

a (n = 25) [Beato et al., 1989]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	5	15	10	**55**	10	**55**	40	15	20	15	0	5	5	0	5
C	25	0	20	10	**75**	25	25	15	35	0	0	0	25	**100**	10
G	**45**	**65**	20	0	0	10	15	25	25	5	**95**	5	5	0	0
T	25	20	**50**	35	15	10	20	45	20	**80**	5	**90**	**65**	0	**85**
Cons	**G**	**G**	**T**	**A**	**C**	**A**	**N**	**N**	**N**	**T**	**G**	**T**	**T**	**C**	**T**

b (n = 81) [GR-TRRD]

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	38	13	44	**66**	6	**72**	26	32	25	8	0	6	6	1	15
C	14	6	17	9	**79**	14	37	13	25	2	2	3	13	**94**	16
G	37	**71**	20	9	10	5	17	24	29	3	**97**	2	8	2	5
T	11	10	19	16	5	9	20	31	21	**87**	1	**89**	**73**	3	**68**
Cons	**G/A**	**G**	**A**	**A**	**C**	**A**	**N**	**N**	**N**	**T**	**G**	**T**	**T**	**C**	**T**

Figure 1. Frequency matrices and GRbss consensus variants obtained by analysis of 25 (*a*) and 81 (*b*) experimentally detected sites. The frequency of nucleotide occurrence is given in %.

An interesting consequence of the analysis of the sample consisting from 152 GRbss is the fact that it contains high percentage of hexameric half-sites participating in glucocorticoid regulation. By analyzing literary data, 3 general mechanisms of action of hexanucleotide GRbss in glucocorticoid induction were supposed. These mechanisms are described below.

1. Functioning of hexameric half-sites as auxiliary elements to the closely located "palindromic" GREs. In regulatory gene regions, hexanucleotide sites are often located in the neighborhoods of the classic GR binding sites organized as inverted repeats. For some of such cases, the presence of hexanucleotide sites is necessary for enhancement (or even for realization) of glucocorticoid response. Functional ensembles of hexanucleotide and "palindromic" GRbss were found in -2,5 kb enhancer of rat tyrosine aminotransferase gene (A00093[4]), promoter regions of human thyrotropin-releasing hormone receptor (A02464), human constitutive androstane receptor (A02491), and rat hepatic aryl sulfotransferase (A02453) genes.

The similar combinations were found also in promoter regions of genes encoding human elastin (A00026), rat angiotensinogen (A00060), rat bone sialoprotein (A00874), rat serine/threonine protein kinase (A00980) and human serine/threonine protein kinase (A02156), enhancer gene regions of rabbit uteroglobin (A00001), rat carbamoylphosphate synthetase 1 (A00757), LTR of Moloney murine sarcoma virus (A00079). It may be supposed that in these genes also, GR-monomers bound to hexanucleotide sites serve as accessory binding factors in addition to GR homodimers bound in the neighborhoods.

2. Interaction (heterodimerization) of GR with the other transcription factors.

As known, affinity for GR-monomer binding to hexameric half-sites is by an order of magnitude lower than affinity for GR-homodimer binding to palindromic sites (Alroy,

[4] Accession No. in TRRD

Freedman, 1992). Hence, on the contrary to "palindromic" sites that are capable to produce glucocorticoid induction of reporter genes even in a single copy, the hexanucleotide site stays inactive in such constructions. Stabilization of binding between GR-monomer and hexanucleotide site may be achieved by formation of the complex (heterodimer) with the other transcription factor, which binds to the neighboring binding site. The best-studied example of such interaction is heterodimerization of GR-monomer with non-related to it protein XGRAF, which binding site closely adjoins the place of binding between receptor and hexanucleotide TGTTCC in promoter region of γ-fibrinogen subunit gene of the clawed frog (A00734). The similar interaction between GR monomer and other proteins takes place in promoter regions of the following genes: rat CYP 27 (Ets2; A01395), mouse α-amylase 2 (PTF1; A00871), mouse glucose-6-phosphotase (FKHR; A00877), in adjacent to promoter regulatory region of the sheep β1-adrenergic receptor (Myc/Max; A01873) gene, and rat atrial natriuretic factor gene (unknown protein; A00954).

3. Strengthening of affinity due to formation of GR multimeres in case hexanucleotide sites are clustered in regulatory gene regions. Clusters participating in glucocorticoid regulation and containing three hexanucleotides located in-between the region 41-88 bp have been found in the following genes: human alcohol dehydrogenase 2, (A00379), mouse phenylalanine hydroxylase (A00768), γ- fibrinogen of clawed frog (A00734), and LTR MMTV (A00045). For one of these clusters, namely TGTTCTgatctgagctcttaTGTTCTattttcctaTGTTCT, in position -120/-80 bp of LTR MMTV, it was shown that affinity of GR to this sequence is the same as to the classic "palindromic" GRE (-191/-167) from the same LTR (Perlmann *et al*., 1990).

Thus, analysis of data accumulated in GR-TRRD has revealed large variability of GRbss structural variants with different mechanisms of glucocorticoid induction.

ACKNOWLEDGEMENTS

The study was partially supported by Siberian Branch of the Russian Academy of Sciences (integration project No. 115). The authors are grateful to I.V. Lokhova for assistance in searching for literature sources.

REFERENCES

Alroy I., Freedman L.P. (1992) DNA binding analysis of glucocorticoid receptor specificity mutants. *Nucl. Acids Res.,* **20**, 1045–1052.

Aranda A., Pascual A. (2001) Nuclear hormone receptors and gene expression. *Phisiol. Rev.,* **81**, 1269–1304.

Beato M., Chalepakis G, Schauer M., Slater E.P. (1989) DNA regulatory elements for steroid hormones. *J. Steroid Biochem.,* **32**, 737–748.

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.

Merkulova T.I., Merkulov V.M., Mitina R.L. (1997) Glucocorticoid regulation mechanisms and glucocorticoid controlled genes regulatory regions: Description in TRRD database. *Mol. Biol. (Russ.)*, **31**, 714–725.

Perlmann T., Eriksson P., Wrange O. (1990) Quntitative analysis of glucocorticoid receptor-DNA interaction at the mouse mammary tumor virus glucocorticoid responsive element. *J. Biol. Chem*., **265**, 17222–17229.

Schoneveld O.J., Gaemers I.C., Lamers W.H. (2004) Mechanisms of glucocorticoid signaling. *Biochim Biophys. Acta*, **1680**, 114–128.

THRESHOLD SELECTION USING THE RANK STATISTICS

Mironov A.A. [1, 2, 3]
[1] Dept. of Bioengineering and Bioinformatics, Moscow State University, Lab. Bldg B, Moscow, 119992, Russia, e-mail mironov@bioinf.fbb.msu.ru; [2] Institite for Information Transmission Problems, RAS, Moscow, 127994, Russia; [3] State Scientific Center GosNIIGenetica, Moscow, 113545, Russia

Key words: threshold, rank statistics, signal search, profile

SUMMARY

Motivation: A common technique in many areas of bioinformatics is calculating a score and comparing this value with a threshold. Usually a training set is used to set the threshold or the threshold is selected *ad hoc*.

Method: This paper describes a natural approach for threshold selection based on rank statistics. Assume a background probability distribution for the considered value and consider a set of n observed scores $\{v_i\}$ that contains a mixture of biologically significant and random values. Sort set of scores by decrease. Using the background probability distribution, calculate probability P_k that at least k out of n observations exceed the observed score v_k. Find $P^* = \min(P_k)$ and let $k^* = \arg\min(P_k)$. We suggest setting the threshold score to v_{k^*}. P^* is the p-value for the selected set of scores. This approach maximizes "non-randomness" of the selected subset of scores. While traditional approaches are based on likelihood or confidence probability that are local and do not take into account the complete data, our approach is global and is based on the analysis of complete dataset.

Results: Applications of this approach to profile construction is presented. In the profile construction problem, the rank statistics technique is applied for selection of significant sequences and selection of significant positions.

Availability: The algorithm for the profile construction is implemented as a WEB server (http://www.bioinf.fbb.msu.ru/SignalX.jsp).

INTRODUCTION

A common technique in many bioinformatics areas is calculating a value (score) and comparing it to a threshold. A typical example is the site search using a profile. In this case the score is calculated at each position of the sequence. If the score at some position exceeds a given threshold, a candidate site is found (e.g. Mironov *et al*., 1999). Another area is the profile construction. Here two problems arise. One is selection of sequences that should be considered, and the other is selection of positions that should form the profile. These problems arise in iterative procedures for profile construction, in particular, such algorithms as MEME (Bailey, Noble, 2003; Kel *et al*., 2004) and the Gibbs sampler (Thompson *et al*., 2003; Favorov *et al*., 2005). This paper describes a natural approach to the threshold selection based on rank statistics. Our approach is based on analysis of entire data set as it is done in (Benjamini, Hochberg, 1995). We describe the general approach and its application to the profile construction.

APPLICATION OF RANK STATISTICS FOR THRESHOLD SELECTION

Let r be a random variable with a known distribution (for example the distribution of profile scores on random sequences): $F(x) = P(r \le x)$. Assume that we have n observations (for example, calculated profile score in n positions of a sequence) producing n values $\{r_i\}$. Fixing a threshold t one selects k_t values. The following question arises: "what is better – to select small number of scores with a high threshold t or to select many scores with moderate t?" To answer this question one can calculate the probability $P_t(k_t)$ that at least k values from the set $\{r_i\}$ exceed the given value t to select the threshold t providing the minimum to the probability $P_t(k_t)$. The motivation here is as follow, biologically significant values (e.g. site scores) should be "nonrandom". We select a threshold t that provides the highest degree of "non-randomness". The obtained minimal value of $P_t(k_t)$ can be used as the p-value for the threshold t.

Clearly it is not necessary to scan all possible threshold values. The best threshold necessarily is one of the observed values r_i. To minimize $P_t(k_t)$, sort the set of observed values by decrease. Then scan the observed values and for every k consider the threshold $t_k = r_k$ and calculate the probability $P_k = P_{t_k}(k)$. The probability p_k that exactly k out of n scores exceed r_k is given by the Bernoulli distribution:

$$pb_k = \mathrm{C}_n^k \, p_k{}^k q_k{}^{n-k}; \qquad p_k = P(r \ge r_k) = 1 - F(r_k), \; q_k = 1 - p_k.$$

The probability P_k that at least k values exceed r_k is obtained by summation of the above probabilities:

$$P_k = \sum_{i=k}^{n} pb_i = \sum_{i=k}^{n} \mathrm{C}_n^i p_k{}^i q_k{}^{n-i}. \tag{1}$$

Scanning over all possible values k produces the minimum for P_k and defines the threshold t:

$$P^* = \min_k P_k; \qquad k^* = \arg\min_k P_k; \qquad t = r_{k^*} \tag{2}$$

The values defined by formulae (1), (2), P_k, P^*, k are random variables because these values are calculated using a set of random variables $\{r_i\}$. It is well known (Balakrishnan, Cohen, 1991) that if $\{r_i\}$ are instances of the same random variable, then the values P_k and k^* have the following properties: for every k P_k is uniformly distributed in the interval [0,1]; k^* is uniformly distributed in the interval [1,n]. These distributions do not depend on the distribution of the source random variable r. The real objects (e.g. the real binding sites) are not random and their scores do not follow the same distribution as the source random variable r. Hence if the data set contains real objects, the values P_k, k^* will not have the above properties of rank statistics and can be used as indicators that allow one to separate the real data with non-random high scores from random noise. The traditional approaches based on likelihoods or confidence probability are local and do not take into account the complete set of data. On the contrary, our approach is global and

involves analysis of the entire dataset. For example, suppose that the dataset $\{r_i\}$ contains 15 values, ten of which have the significance $P_k = 0.1$. In this case the local approach provides week significance 0.1 while our approach produces the significance $1.9 \cdot 10^{-7}$. On the other hand, if only one observation out of fifteen has significance 0.1 our approach will give significance of the dataset 0.79.

SELECTION OF SIGNIFICANT POSITIONS IN MULTIPLE ALIGNMENT

Recognition profile constructed based on a multiple alignment should include only significant positions; otherwise, the output would be overwhelmed by noise. The information content I_k is a natural quality measure for an alignment column k :

$$I_k = \sum_{\alpha} f_{\alpha}^{k} \log\left(\frac{f_{\alpha}^{k}}{f_{\alpha}}\right). \tag{3}$$

Here f_{α}^{k} is the observed frequency character α in column k, f_{α} is the background frequency. The columns with low information content are insignificant, whereas the columns with high information content are significant, so the problem is to set the threshold for separation of significant and insignificant positions. To apply the rank statistic technique for threshold selection one need the background probability distribution of the information content. This distribution can be calculated in some simple cases (for example if the character probabilities are uniform), but in the general case it is unknown. We can avoid this problem if we assume the type of the probability distribution for I . For example, we can assume that the distribution for I is normal or exponential. We have applied this approach to the threshold selection in the problem of identification of specificity determining positions in protein alignments (Kalinina *et al*., 2004).

SELECTION OF SEQUENCES FOR PROFILE CONSTRUCTION

A typical problem of signal identification is as follows. Given a set of sequences that presumably contain sites representing a signal, find this signal. The most popular algorithms addressing this problem are MEME (Bailey, Noble, 2003; Kel *et al*., 2004) and the Gibbs sampler (Thompson *et al*., 2003; Favorov *et al*., 2005). In real biological situations there is no guarantee that all sequences contain sites. Methods of comparative genomics can provide up to 50 % of sequences that may not contain sites, whereas expression arrays often produce even more irrelevant sequences.

Here we describe application of the rank statistics to select appropriate (site-containing) sequences in the MEME setting. The algorithm identifies the highest-scoring hits of the current profile in each sequence. Then using the selected sites, it reconstructs the profile. At that point our technique can be used to retrain only significant sites. We use the current profile score as a measure of the site quality. We assume that the profile score has the normal distribution (profile score is sum of independent random variables, positional nucleotide weights). Hence, the probability that the score exceeds a given value can be determined. Using the rank statistics, we can select sites that should be included in the profile. The modified MEME algorithm will have the following iterations:

1. Select site and create a profile.
2. Iterate:
    a. Find the best hit of the current profile in each sequence.

b. Sort sites by score.
c. Using rank statistics define a threshold and select a subset of significant sites.
d. Using selected sites, create the new profile.
e. Using rank statistics select positions that should be included in the profile.

DISCUSSION

This approach can be applied to different tasks of bioinformatics, in particular for regulation analysis in genomes. This technique allows one to select a non-random subset of observations against a background of random observations. This approach is not completely free from risks and problems. Firstly, the set of observed values should contain a significant number of random events. Otherwise the minimum of the P_k may select a subset with a very strong threshold, e.g. the selection may contain only one observation. On the other hand even the observations contain only random values from a given distribution, P_k will have a minimum at some position. This value is a uniformly distributed random variable. If P_k are independent, the significance of p-value (minimum of P_k) can be evaluated using the extreme value distribution: $P(\text{p - value} < x) = P\left(\min(P_k) < x\right) = 1 - \left(1 - F(x)\right)^n \cong nx$. But the random variables P_k are not independent and thus this is an incorrect estimation. The values P_k are correlated and this is too pessimistic. Thus if we see a p-value considerably less than $1/n$, then we have obtained a significant subset. In applications, a detailed investigation of the p-value behavior may lead to interesting observations. A secondary minimum may become the main minimum if one changes the scoring scheme or the definition of the background probability. On the other hand, deep secondary minima may be biologically reasonable.

ACKNOWLEDGEMENTS

I am grateful to Mikhail Gelfand for fruitful discussions, careful reading of the manuscript, and helpful remarks. This study was partially supported by Grants from the Russian Academy of Sciences (programs “Molecular and Cellular Biology” and “Origin and Evolution of Biosphere”), the Howard Hughes Medical Institute (grant No. 55000309), and the Russian Fund of Basic Research (grants Nos 04-04-49438, 06-01-00454).

REFERENCES

Bailey T.L., Noble W.S. (2003) Searching for statistically significant regulatory modules. *Bioinformatics*, **19**, Suppl 2, II16–II25.

Balakrishnan N., Cohen A. C. (1991) *Order Statistics and Inference*. Academic Press, New York, 7–17.

Benjamini Y., Hochberg Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Stat. Soc. B*., **57**, 289–300.

Favorov A.V. *et al*. (2005) A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics*, **21**, 2240–2245.

Kalinina O.V. *et al*. (2004) Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families. *Protein Sci.*, **13**, 443–456.

Kel A. *et al*. (2004) Recognition of multiple patterns in unaligned sets of sequences: comparison of kernel clustering method with other methods. *Bioinformatics*, **20**, 1512–1516.

Mironov A.A. *et al*. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucl. Acids Res*., **27**, 2981–2989.

Thompson W. *et al*. (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucl. Acids Res*., **31**, 3580–3585.

STUDIES ON TRANSCRIPTIONAL REGULATION IN DNA

Mitra Ch.K.
School of Life Sciences, University of Hyderabad, Hyderabad, 500 046, India,
e-mail: c_mitra@yahoo.com

Key words: core promoter; transcription sites; information content; substitution matrices

SUMMARY

Motivation: We have studied the core promoter regions and also the transcription factor binding sites (TFBS) in DNA using the information content of substitution matrices. The core promoter region and the transcription factor binding site databases have been obtained from the Internet and used without modification. The databases are already aligned for direct use. We have determined the substitution matrices of core promoter by the direct counting method over a block size of 5, 11 and 15 nucleotides. Information content has been directly plotted from these matrices in the form of histograms. Similar approaches have been utilized for the transcription factor binding sites except that we have used the JASPAR database and have used the DRAGRAM of PHYLIP package to show our results as a tree diagram.

Results: We notice that the information content peaks in around the 11 nucleotide region around the TSS. We also note that several of the transcription sites are very similar, as determined by the phylogenic studies. The 11 nucleotide range is optimal, as a 5 nucleotide range may be too frequent (1 in 2^5) and the 15 nucleotide sequence may be too infrequent (1 in 2^{15}). It is also possible and very likely that the binding may not be sequence dependent but depends only on the local conformation.

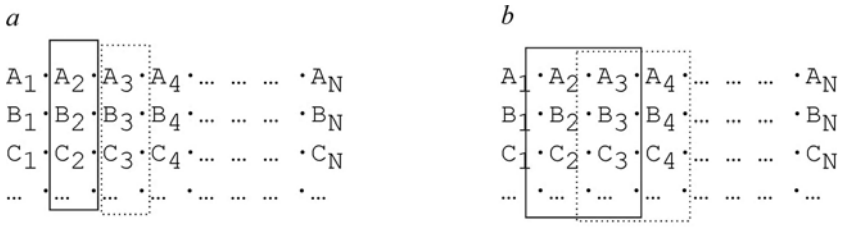
INTRODUCTION

The recent genome projects revealed that in eukaryotes the coding region is very less than expected before. Human genome contains less than 25,000 genes that represent less than 2 % of the whole genome. Unlike in most prokaryotic genomes that contain packed gene units with few intergenic regions, repeated and non-coding sequences that do not code for proteins make up the remaining part of the human genome. Gene expression and its regulation involve the binding of many regulatory transcription factors (TFs) to specific DNA elements called Transcription Factor Binding Sites (TFBS). Promoter region is a regulatory region of the protein-coding genes and shows variation from species to species. The transcription factors (cell or tissue specific) bind to the promoter region of the DNA that subsequently causes efficient binding of RNA polymerase to initiate mRNA synthesis. Specific DNA sequence elements within the promoter region (like TATA-box, CCAAT-box, Downstream Promoter Element (DPE) and GC-box) exhibit similarities between different promoters of the same DNA as well as between various species. The core promoter region (which can extend ~35 bp upstream and which is a minimal promoter region required to start the pre-initiation complex formation) usually has TATA-box, which is conserved in most of the species (30–50 % of promoters) and TSS region, which usually is not conserved. Each nucleotide in the consensus sequence motif (TATA box, CCAAT box and GC box) represents the most frequently occurring nucleotide at that position and does not represent an actual sequence. Reliable

identification of the core promoter region by RNA polymerase II prior to transcription initiation is mandatory for the proper initiation and regulation of mRNA synthesis. The region 200–300 bp immediately upstream of the core promoter is the proximal promoter that has abundant of TFBS. Further upstream is the distal promoter region that usually contains enhancers and few TFBS. TFBS are represented by relatively short (5–10 bp) nucleotide sequences. Specificity of TF is defined by its interaction with TFBS and it is extremely selective, mediated by non-covalent interactions between appropriately arranged structural motifs of the TF and exposed surfaces of the DNA bases and backbone The ability of the cell to control the expression of genes under different developmental and environmental conditions is still poorly understood. Identifying functional TFBS is a difficult task because most TFBS are short, degenerate sequences occurring frequently in the genome. The non-coding sequences play a crucial role in gene regulation hence the computational identification and characterization of these regions is very important.

METHODS

The counting process needed for the mono and dinucleotide substitution matrices can be easily seen from the following diagram:



The principle of counting the frequencies illustrated diagrammatically. The left side diagram (*a*) shows the counting principle for neighbor-independent frequency determination. The three lines show the nucleic acid bases corresponding to the TFBS already aligned in the database. The solid box is used for determination of the actual frequencies and the counts for A_2-B_2, A_2-C_2 and B_2-C_2 are put in a 4 × 4 matrix. Then the counting box is shifted by one position (dotted box) and the process is repeated. In the right side illustration (*b*), we indicate the counting principle for neighbor-dependent (pair-wise) determination of frequencies. In this illustration, we get the actual counts for A_2A_3-B_2B_3, A_2A_3-C_2C_3, B_2B_3-C_2C_3 and these are placed in a 16 × 16 matrix. The counting box is next moved right by one base position (shown by the dotted box) and the process continued till the TFBS region is completed.

The information content of the substitution matrices have been computed using the classical formula: $H = \sum_{ij} q_{ij} s_{ij} = \sum_{ij} q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}$. Where H is the information content, q_{ij}'s are the observed frequencies and s_{ij}'s are the elements of the substitution matrix. For the dinucleotide variant of this computation, the ideas remain the same (we shall be having four subscripts grouped in two pairs) (Altschul, 1991).

RESULTS

We have plotted the information content of the core promoter region for the human and mouse genome (Périe *et al*., 1998) for the three regions indicated in Fig. 1. We notice

that near the core promoter region, a blocksize of 5 nucleotides give a strong signal whereas for larger blocksizes, the signal drops off rapidly, as expected (Reddy *et al*., 2006a). (We have also studied the plant (Shahmuradov *et al*., 2003) and *E. coli* (Hershberg *et al*., 2001) promoters but the data is not shown here).

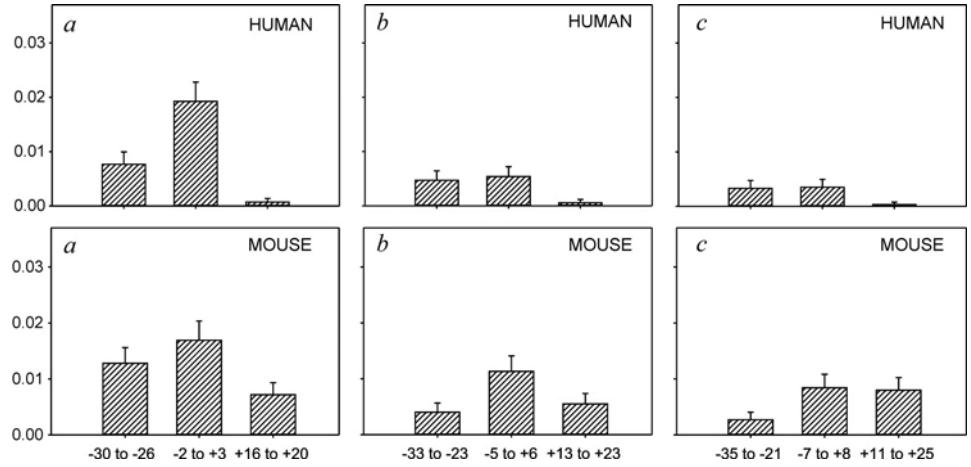


Figure 1. The average mutual information content H, (in bits) of core promoter elements (calculated by neighbor-independent nucleotide substitutions) from different datasets. In all the figures "a", "b" and "c" represents block size 5,11 and 15 respectively. Each graph has three bars representing TATA-box region, TSS region and downstream region. The bars on top of the histograms represent the standard errors of the 16 H_{ij} values.

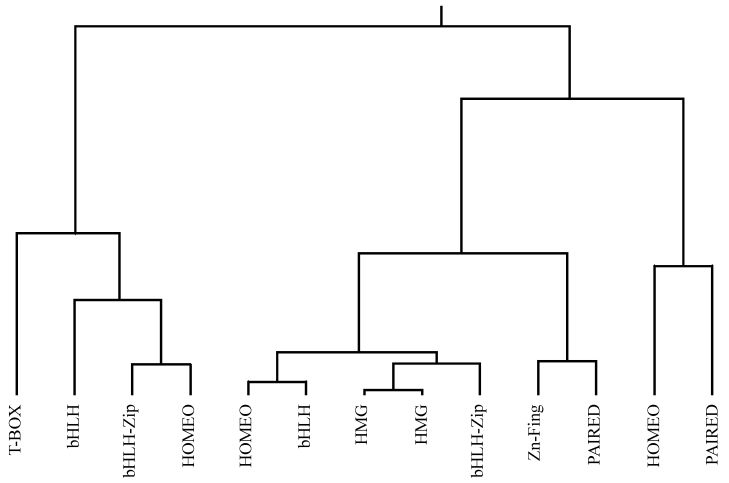


Figure 2. Functional classification of TFBS in mouse; information content is calculated from nucleotide for the neighbor-dependent substitution matrices. We have indicated the TF's by their class names as this may help to see the relation between them.

In Fig. 2 we have plotted the TFBS information content as a tree diagram (using the PHYLIP package) for the mouse genome, as obtained from the JASPAR database (Sandelin *et al*., 2004). This information content has been obtained from the neighbor dependent computations (i.e., a 16×16 substitution matrix has been used). Again we note that a group of related binding sites are getting clustered suggesting a common transcription for these proteins. These may have a strong functional implication (Reddy *et al*., 2006b).

DISCUSSION

We know that the core promoter region and the transcription factor binding sites play important roles in the transcription process that is prior to the translation step. The transcription process is governed by several factors (or proteins) that must initiate the production of mRNA from the DNA. Although a lot is known about the translation mechanism, relatively less is known about the transcription process, in particular the detailed ideas about the factors that cause the initiation of transcription. We note that (i) there are several sites that are quite similar so as to bind a given transcription factor and (ii) the core promoter region is essentially small and is expected to be around 11 bases in size.

ACKNOWLEDGEMENTS

We gratefully acknowledge financial support from various funding bodies that made this work possible.

REFERENCES

Altschul S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol*., **219**, 555–565.

Hershberg R. *et al*. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucl. Acids Res*., **29**, 277.

Karlin S., Altschul S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, **87,** 2264–2268.

Périe C.R. *et al*. (1998) The eukaryotic promoter database EPD. *Nucl. Acids Res*., **26**, 353–357.

Reddy D.A. *et al*. (2006a) Comparative analysis of core promoter region: information content from mono and dinucleotide substitution matrices. *Computational Biology and Chemistry*, **30**, 58–62.

Reddy D.A. *et al*. (2006b) Functional classification of Transcription Factor Binding Sites: Information content as a Metric. *J. of Integrative Bioinformatics*, 0020 (Online Journal).

Sandelin A. *et al*. (2004) JASPAR: an open access database for eukaryotic transcription factor binding profiles. *Nucl. Acids Res*., **32**(1), D91–D94.

Shahmuradov I.A. *et al*. (2003) PlantProm: a database of plant promoter sequences. *Nucl. Acids Res*., **31**, 114–117.

MODELING TRANSCRIPTIONAL REGULATION WITH EQUILIBRIUM MOLECULAR COMPLEX COMPOSITION

Mjolsness E.
Institute for Genomics and Bioinformatics, and Departments of Computer Science and Mathematics
University of California, Irvine, CA, USA
Corresponding author: e-mail: emj@uci.edu

Key words: equilibrium, composition, EMCC, cooperative activation, statistical mechanics, partition function, allosteric enzyme, transcriptional regulation, systems biology

SUMMARY

Motivation: Regulation of transcription has been modeled in a variety of ways in cellular and developmental systems.

Results: Here we apply a method for creating equilibrium models of hierarchical statistical systems, the Equilibrium Molecular Complex Composition (EMCC) family of models, to the problem of modeling the rate of initiation of transcription in the presence of overlapping binding sites, synergistic binding interactions in one dimension, and modular activation of a transcription complex.

INTRODUCTION

The essential steps for modeling a hierarchical system in equilibrium using the Equilibrium Molecular Complex Composition (EMCC) family of models are to (1) identify the hierarchical levels; (2) model each level with a partition function Z for a Boltzmann distribution, as a function of fugacity parameters $\mathbf{z}$ for constituent molecules or subcomplexes; (3) perform any possible model reduction (including justifiable approximations) on the resulting partition functions $Z(\mathbf{z})$; (4) compose the partition functions, substituting partition functions Z from a finer scale for fugacities $\mathbf{z}$ at a coarser scale. The validity of this procedure follows from an EMCC "Composition Theorem". We will illustrate this procedure in the case of the Monod Wyman Changeaux model of allosteric enzymes, and then apply it to the case of a hierarchical model of transcriptional regulation (Mjolsness, 2001) here generalized to the case of transcription factor binding sites with optional overlaps with their nearest neighbors in one dimension, and optional interaction energies with their second nearest neighbors, and hierarchical activation in terms of transcriptional regulatory modules.

Assume we have a molecular complex defined at each level by a set of binary occupancy variables $s_i \in \{0,1\}$, related through a high-order Ising model. For each slot there is a fugacity variable z_i. We can define a multidimensional array J of interaction energies, whose elements are indexed by the ordered set of indices $\rho(\sigma)$:

$$J_{\rho(\sigma)} = J_{(i(1)<i(2)<\ldots i(l))} \in \mathbb{R}$$

with the convention that any other values of J are 0. Defining $0^0 = 1$, the partition function for equilibrium statistical mechanics is

$$Z(z \mid J) = \sum_{\{s \mid s_i \in \{0,1\}\}} (\prod_i z_i^{S_i}) \prod_{\{\sigma \mid \sigma_i \in \{0,1\}\}} \exp[-\beta J_{\rho(\sigma)} \prod_j (s_j)^{\sigma_j}] \quad (1)$$

Considered as a function of the fugacities $\mathbf{z}$, $Z(\mathbf{z})$ is a high-order polynomial and it is a generating function for the (unnormalized) probabilities of all configurations s. However, many J's can tend towards ∞ in such a way as to prohibit particular combinations of values of s_i by giving them zero probability. Also many J's can be exactly zero, so that particular interactions are absent. These possibilities can be encoded by the predicates $P(s)$ and $Q(\sigma)$, respectively, in the following expression for the partition function:

$$\begin{aligned} Z(z \mid J) &= \sum_{\{s \mid P(s)\}} (\prod_i z_i^{s_i}) \prod_{\{\sigma \mid Q(\sigma)\}} \exp[-\beta J_{\rho(\sigma)} \prod_j (s_j)^{\sigma_j}] \\ &\equiv \sum_{\{s \mid P(s)\}} (\prod_i z_i^{s_i}) \prod_{\{\sigma \mid Q(\sigma) \wedge (\wedge_i (\sigma_i \Rightarrow s_i))\}} (\omega)_{\rho(\sigma)} \end{aligned} \quad (2)$$

As a trivial example, a heterodimer of species 1 and 2 with no internal states would have $Z(z_1, z_2) = \omega_{1,2} z_1 z_2$. A protein with a single binding site that can be empty or occupied by species 1 or 2 would have $Z(z_1, z_2) = 1+\omega_1 z_1+\omega_2 z_2$. If the protein is itself regarded as one of the species that can be present or absent, with fugacity z_0, then it must be present and the partition function is $Z(z_1, z_2) = z_0(1+\omega_1 z_1+\omega_2 z_2)$. In each case, as for any probability generating function, the coefficients can be normalized to give the probabilities of each possible configuration of bindings.

Such partition functions can be put into a form with homogeneous degree by introducing the complementary fugacity variables $z_i = z_i^+ z_i^-$: $Z^{\text{homog}}(z^+, z^- \mid \omega) = Z^{\text{homog}}(z^+ / z^- \mid \omega)(\prod_i z_i^-)$. No information is lost since $Z^{\text{homog}}(z \mid \omega) = Z^{\text{homog}}(z^+ = z, z^- = 1 \mid \omega)$.

METHODS AND ALGORITHMS

Composition Theorem. Suppose we have a two-level hierarchical system, with a top level (coarse-scale) partition function Z_0 and a set of lower-level (finer-scale) partition functions. Given partition top-level internal state variables $\{s_0\}$ that can interact with lower-level systems, and lower-level activation variables p_i that can interact with higher-level systems, we can define lower-level partition functions $Z_i^{([s_{0j}],p_i)}(z, \omega)$. Without the indices s_0 and p_i, generating functions for discrete-time branching processes (birth-and-death processes) are obtained by function composition from the generating functions at each succeeding generation, with the first generation as the outermost composition (Athreyea, Ney, 1972). A similar result holds in the present situation.

A "Composition Theorem" gives conditions under which partition functions $Z_0(\mathbf{z})$ at the top level and $\{Z_i(\mathbf{z}_i) \mid i \geq 1\}$ at the next lower level in a scale hierarchy, all of which are in the form of (Equation 2), may be *composed* to give the partition function $Z_{2-\text{level}}([\zeta_i Z_i(\mathbf{z}_i) \mid i \geq 1])$, also in the form of (Equation 2), for the composite molecular complex. Optionally some of the ζ_i may be set to 1 if we do not need to differentiate

with respect to them. For example, if $Z_i(\boldsymbol{z}_i) = (z_{1i})^2$ then there is a model level corresponding to obligatory homodimerization in binding to the top-level complex at position *i*. Likewise if $Z_i(\boldsymbol{z}_i) = (1 + \omega_1 z_{1i} + \omega_2 z_{2i})$, then there is a binding site which can be empty, or occupied by just one of two competing factors. The composition theorem may be used recursively to model many levels of complex composition.

IMPLEMENTATION AND RESULTS

MWC Example. A simple example is given by the Monod-Wyman-Changeaux model of allosteric enzymes.

Level 1 (top): global activation/inactivation: $Z^1 = \zeta_i \omega_0 Z^{2+} + Z^{2-}$.

Level 2: Independent identical subunits: $Z^{2\pm} = (Z^{3\pm})^n$.

Note: levels 1 and 2 are ordinarily combined.

Level 3: Independent binding heterogeneous sites within each subunit: $Z^{3\pm} = (\prod_{\alpha=1}^{A} Z_\alpha^{4\pm})^n$. The simplest case is $\alpha \in \{1, 2, 3\}$ for substrate/product, activator, and inhibitor respectively.

Level 4: Mutual exclusion (MutEx) for occupation: $Z_\alpha^{4\pm} = \omega_\alpha^{\pm} + \sum_{i=1}^{n} \omega_{\alpha i}^{\pm} Z_{\alpha i}^{5\pm}$.

Without loss of generality, take $\omega_\alpha^{\pm}$=1 since empty binding sites are never prohibited.

Level 5: Convergence through sharing of fugacity variables, each of which is (for a dilute well-stirred solution in a fixed macroscopic volume) proportional to the number of molecules present and therefore to concentration: $Z_\alpha^{5\pm} = z_i$.

Composition of all levels: $Z = z_0 \omega_0 \prod_{\alpha=1}^{A} (1 + \sum_{i=1}^{n} \omega_{\alpha i}^{+} z_i)^n + \prod_{\alpha=1}^{A} (1 + \sum_{i=1}^{n} \omega_{\alpha i}^{-} z_i)^n$.

The original MWC model has $\omega_{\alpha i}^{\pm} = 0$ unless $I = \alpha$ and the following condition: $(s = +1 \wedge (\alpha = 1 \vee \alpha = 2)) \vee (s = -1 \wedge (\alpha = 1 \vee \alpha = 3))$ where α = (1, 2, 3) for substrate, activator, and inhibitor respectively. In that case we recover the original MWC model:

$$\begin{aligned} Z = {} & L(1 + \textstyle\sum_{i=1}^{n} c(S/K_S))^n (1 + \sum_{i=1}^{n} (A/K_A))^n \\ & + (1 + \textstyle\sum_{i=1}^{n} c(S/K_S))^n (1 + \sum_{i=1}^{n} (I/K_I))^n \end{aligned} \tag{3}$$

Clearly this model can be generalized to multiple substrates, activators and inhibitors on each subunit, as demonstrated and applied in (Tarek *et al*., 2006) to amino acid synthesis pathways.

EMCC application to transcriptional regulation. With this apparatus we can rederive and extend a model similar to Hierarchical Cooperative Activation (Mjolsness, 2001) for transcriptional regulation. Transcription factors bind, alone or in multimers such as homodimers or heterodimers, to DNA binding sites that can overlap with their one-dimensional neighbors (in which case they can't be occupied simultaneously) or be sufficient close to their nearest nonoverlapping neighboring sites in one dimension that energetic interactions occur. These possibilities are summarized by allowing overlap with nearest neighbors to either side, interaction with next nearest neighbors to either side, and missing sites that break chains of overlap and/or interaction. At a coarser level, activation occurs in modules or cassettes (such as the Drosophila even-skipped minimal stripe three element) which contribute to overall activation of transcriptional initiation. Within these limitations, we can formulate an equilibrium complex model similar to MWC at several levels. The novel part of this model compared to HCA is the one-dimensional interactions

through site overlap and synergy: second nearest neighbors (odd or even) interact energetically with factor ω. Therefore each successive pair of sites has three possible states. The model can be solved using 3 × 3 transfer matrices on site pairs:

$$Z = (1,1,1)\cdot\{\prod_{i=k\searrow 1}\begin{pmatrix} 1 & 1 & 1 \\ z_{2i+1} & z_{2i+1}\omega_{2i-1,2i+1} & 0 \\ z_{2i+2} & z_{2i+2} & z_{2i+2}\omega_{2i,2i+2} \end{pmatrix}\}\cdot\begin{pmatrix} 1 \\ z_1 \\ z_2 \end{pmatrix}$$

Any site can be omitted (removing its overlap constraints and interaction energies) by setting its z_i to be 1 and $\omega_{i*} = \omega_{*I} = 1$.

ACKNOWLEDGEMENTS

Thanks to Vitali Likhoshvai, Bruce Shapiro, Tarek Najdi, and Chin-Ran Yang for valuable discussions. This work was supported by Biomedical Information Science and Technology Initiative (BISTI) grant (No. 4 R33 GM069013) from the National Institute of General Medical Sciences, USA, and by US National Science Foundation Frontiers in Biological Research (FIBR) grant No. EF-0330786.

REFERENCES

Athreyea K.B., Ney P.E. (1972) Branching Processes. Springer-Verlag; Dover.

Mjolsness E.D. (2001) Gene Regulation Networks for Modeling Drosophila Development. In Bower J.M., Bolouri H., (eds), *Computational Methods in Molecular Biology*, MIT Press.

Tarek S.N., Chin-Ran Yang, Shapiro B.E., Hatfield G.W., Mjolsness E.D. (2006) Application of a Generalized MWC Model for the Mathematical Simulation of Metabolic Pathways Regulated by Allosteric Enzymes. *J. of Bioinformatics and Computat. Biol*. to appear.

OWEN-SCRIPT – AN EXTENDED TOOL FOR PAIRWISE GENOME ALIGNMENT

***Ogurtsov A.Yu.*[*1], *Vasilchenko A.N.*[2], *Vlasov P.K.*[3], *Shabalina S.A.*[1], *Kondrashov A.S.*[1], *Roytberg M.A.*[*2]**

[1] National Center for Biotechnology Information, NIH, 45 Center Drive, Bethesda, MD 20892-6510, USA; [2] Institute of Mathematical Problems in Biology, Pushchino, Moscow Region, 142290, Russia; [3] Institute of Molecular Biology, Moscow, 117234, Russia

[*] Corresponding authors: e-mail: Ogurtsov@ncbi.nlm.nih.gov, Roytberg@impb.psn.ru

Key words: pairwise alignment, genomes, hierarchical approach

SUMMARY

Motivation: A genome alignment is an important instrument of post-genomic computational biology. The commonly available tools (LAGAN, BLAT, YASS, etc) designed for command line mode and thus tend to loose some similarities without any possibility for user to learn about this. In contrast, the OWEN is an interactive tool, allowing user to control the alignment process and to be sure that no interesting events were lost. However, one may need tools to store some alignment protocols that are suitable for a class of similar situations and then implement the protocol automatically.

Results: We propose OWEN-SCRIPT, an extension of the OWEN program thet al.lows to perform OWEN based scripts. The commands of the scripts correspond to the actions of interactive OWEN. Examples of protocols obtained from alignment human and mouse genomes are also available.

Availability: Program OWEN-SCRIPT is available on request from the authors.

INTRODUCTION

OWEN, named after a scientist who developed the concept of homology (Owen, 1848) is a software tool for aligning pairs of long sequences based on greedy paradigm (Roytberg *et al*., 2002). Unlike other popular tools (e.g. LAGAN (Brudno *et al*., 2003), YASS (Noe, Kucherov, 2005), etc.) OWEN is an interactive tool and allows human intervention at every step of the alignment process. This makes the user sure that (s)he did not miss any essential similarity. Constructing a detailed alignment usually takes 5–15 iterative steps; each of steps consists of constructing and editing local similarities and with resolving conflicts between them.

However an alignment protocols invented during the interactive work can be adequate for a series similar cases. Because of this we have implemented in the OWEN a script option. The script commands are in almost one-to-one correspondence with the interactive actions and thus any alignment protocol can be represented with a proper script; the script then can be used to align automatically proper genome pairs.

METHODS AND ALGORITHMS

OWEN actions: an overview. OWEN session starts with the determination of input data. During the session OWEN stores a set of local alignments. All alignments can be divided in two classes: those that are in conflict with some other alignments, and the non-conflicting ones, i.e. those are collinear to any other alignment (two alignments are collinear if segments involved in one of them precede in both sequences segments involved in the other, and are in conflict otherwise). The aligning with OWEN consists mainly of creating, editing, and deleting local alignments; the corresponding actions are listed under CONSTRUCT, CONFLICT and FILTER items of the main menu.

Actions listed under CONSTRUCT create new and modify the already present alignments, e.g. *Align* creates new alignments in areas defined by the present alignments (so that the new alignment cannot be in conflict with any of the present ones); *Expand* extends existing alignments. Actions listed under CONFLICT resolve conflicts between alignments by trimming conflicting alignments (*Reconcile*) or by completely deleting some of them (*Greedy, Optimal*, and *Kill*). Actions listed under FILTER can create, update, and delete the filter. A filter is a list of segments in both sequences that (i) are annotated as repeats, and/or (ii) are aligned with several segments in the other sequence, and/or (iii) have low complexity. Segments included in the filter can be masked when actions *Align* and *Expand* are performed.

The ultimate goal of a session is to construct the best (from the user's point of view) chain of non-conflicting alignments, then to fill the gaps between them by the algorithm of global sequence alignment and thus to obtain the global alignment of the given sequences. However, the user can produce and save different global alignment and/or save intermediate sets of local alignments that possibly contain conflicts.

IMPLEMENTATION

OWEN command file. The OWEN command file is a text file, each it's line is an operator, describing an action of OWEN and parameters of the action. For example, sequence1 chr6_hum.seq 1–1000000 causes an input of first million of nucleotides from a file chr6_hum.seq as the first sequence of the pair to be aligned. The operator align p = 0.000001 w = 16 5/8 = 12 nomask leads to generation of all local similarities, which have P-value below p = 0.000001, and are detectable with given parameters of the algorithm, e.g. they should contain at least 16 consecutive matches. The sequences to be analyzed should be prepared by preceding operators. All OWEN actions can be represented with proper operators. The complete list of operators (commands) and corresponding actions is given in the Manual, available at ftp://ftp.ncbi.nih.gov/pub/kondrashov/owen. The web site also contains templates of command files; using the templates one can create command files adjusted to a typical biological problems. Protocols of genome alignments and their script representations are also discussed in (Ogurtsov, 2005).

The operators of OWEN-SCRIPT command file are performed one by one, condition operators are not allowed. We have declined implementation of BASIC-like command language, because the developed tool was sufficient to solve all problems arisen in our work.

The general form of the OWEN command file is given on Fig. 1.

The command file *owen.cmd* can be executed with the command>owen owen.cmd Scripts based on OWEN command files.

A simple, but important way to extend abilities to describe alignment tasks is to utilize UNIX scripts, MS WINDOWS batch files, or analogous resources of other operating systems. This is the way, for example, to prepare a task to align a large set of sequence pairs.

```
// INPUT
start hide
sequence1 HUMAN.seq  0 - 0
sequence2 MOUSE.seq  0 - 0 invcompl

……..

// OUTPUT of RESULTS
save Hum_Mus.gal as global
exit
```

Figure 1. General form of the OWEN command file. The file determines alignment of the sequences from files HUMAN.seq and MOUSE.seq; the latter should be invert-complemented. Results will be stored in the file Hum_Mus.gal.

Indeed, having an OWEN command file (see Fig. 1), describing an alignment protocol, one can easily create a UNIX script (see Fig. 2) that provides an alignment of given sequence according the protocol. To obtain the script one needs (1) add "echo" at the beginning of each line of the command file; (2) add ">> owencmd.tmp" at the end of each line or "> owencmd.tmp" at the end of the 1st line; (3) add two lines at the bottom of the file:

owen owencmd.tmp

delete owencmd.tmp

The obtained script will create a file owencmd.tmp, which is a copy of an initial command file, then run OWEN with the command file and delete command file. By substitution of any parameter of the script with "$1", "$2", etc. one can obtain a parameterized script (see Fig. 2).

```
// INPUT
echo start hide                          >owencmd.tmp
echo sequence1 $1  0 - 0                >>owencmd.tmp
echo sequence2  $2  0 - 0 invcompl           >>owencmd.tmp

……..

// OUTPUT of RESULTS
echo save $3 as global                            >>owencmd.tmp
echo exit                                       >>owencmd.tmp
owen owencmd.tmp
delete owencmd.tmp
```

Figure 2. UNIX script obtained from the command file given on Fig. 1. The names of input and output files are described as parameters of the script.

For example, suppose, that the file Align-1.sh contains a copy of the script from Fig. 2. Then for any files seq_A.txt and seq_B.txt the script Align-1.sh seq_A.txt seq_B.txt Result_AB will provide the alignment of the sequences from the files according the protocol of Fig. 1 and output results to the file Result_AB. The script Align-1.sh, in turn, can be called from another script, etc. Examples of scripts can be found at ftp://ftp.ncbi.nih.gov/pub/kondrashov/owen.

Basic tools, environment and architecture. The OWEN-SCRIPT's source is portable. It is written on ANSI C++, the total volume is ~ 10 000 lines. Graphic interface is based on the Fox-toolkit (see http://www.fox-toolkit.org/). All libraries are linked as static, this

guaranties that executable module can be downloaded and run *per ce* on user's computer with the same processor type.

OWEN's architecture can be represented as a finite automaton. Receiving an input signal (user's click in interactive mode or command line in a batch mode), it performs a corresponding action. The list of actions is given in the Manual. The data structures are mainly same as in previous version of OWEN (Ogurtsov *et al*., 2002). The main data type is *a box*, i.e. a pair of fragment U[a1, a2] and V[b1, b2] of given sequences U and V. For each box we remember a non-conflicting chain of local similarities ("backbone chain", see (Roytberg *et al*., 2002)), its score and some other values. The boxes are arranged in 3 trees, which support quick search by both coordinates of a block and its score.

CONCLUSION

OWEN-SCRIPT is a powerful tool and have been used in many works (see e.g. (Bazykin *et al*., 2004; Ogurtsov *et al*., 2004; Shabalina *et al*., 2004). The main advantage of the tool is its ability to fit the specificity of the data and then reproduce the obtained procedure of analysis.

ACKNOWLEDGEMENTS

This work was supported by the Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS. Roytberg M.A. acknowledges financial support by the Russian Foundation for Basic Research (project Nos 03-04-49469, 02-07-90412), by grant from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003), NWO, ECO-NET, NIH (grant TW005899, co-PIs V. Tumanyan and M. Borodovsky) and by the program of RF Ministry of Science and Education (contract No. 02.434.11.1008).

REFERENCES

Bazykin G.A., Kondrashov F.A., Ogurtsov A.Y., Sunyaev S., Kondrashov A.S. (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. *Nature*, **429**(6991), 558–562.

Brudno M., Do C.B., Cooper G.M., Kim M.F., Davydov E., Green E.D., Sidow A., Batzoglou S. (2003) NISC Comparative Sequencing Program. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res*. **13**(4), 721–731.

Noe L., Kucherov G. (2005) YASS: enhancing the sensitivity of DNA similarity search. *Nucl. Acids Res.*, **33**(Web Server issue), W540–543.

Ogurtsov A.Iu. (2005) A protocol of automatic alignment of genome sequences using the program OWEN. *Biofizika*, **50**(3), 475–479. (In Russ.).

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

Ogurtsov A.Y., Sunyaev S., Kondrashov A.S. (2004) Indel-based evolutionary distance and mouse-human divergence. *Genome Res.*, **14**(8), 1610–1616.

Owen R. (1848) On the archetype and homologies of the vertebrate skeleton. London, John Van Voorst.

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

Shabalina S.A, Ogurtsov A.Y., Rogozin I.B., Koonin E.V., Lipman D.J. (2004) Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucl. Acids Res*., **32**(5), 1774–1782.

A COMPREHENSIVE QUALITY ASSESSMENT OF THE AFFYMETRIX U133A&B PROBESETS BY AN INTEGRATIVE GENOMIC AND CLINICAL DATA ANALYSIS APPROACH

Orlov Yu.L.[1], Zhou J.T.[1], Lipovich L.[1], Yong H.C.[2], Li Yi[2], Shahab A.[2], Kuznetsov V.A.[*1]
[1] Genome Institute of Singapore, Singapore; [2] Bioinformatics Institute, Singapore
[*] Corresponding author: e-mails: orlovy@gis.a-star.edu.sg; kuznetsov@gis.a-star.edu.sg

Key words: Affymetrix, database, human genome, repeats, anti-sense, clinical data analysis, breast cancer

SUMMARY

Motivation: Insufficient reliability of expression measurements is key problem facing microarray experiments. The problem could originate from poor gene identification by the probe sequences, whose design may not consider the actual complexity of the human genome.

Results: We re-estimated genome localization of the Affymetrix U133A and U133B GeneChip (initial) target sequences. We matched these sequences to gene and transcripts in the human genome. This resulted in the significant redefinition of specificity and uniqueness of more than 2500 GeneChip probesets. Among the rest target sequences, approximately one quarter overlapped with interspersed repeats that could cause cross-hybridization signals and errors in expression measurements. To test that hypothesis, we compared GeneChip microarray data from large groups of breast cancer patients differed by aggressiveness of tumor growth. In particular, for low- and high- aggressive tumors, we demonstrated that among the set of differentially expressed genes the probesets with of repeat-overlapped target sequences statistically significant underrepresented in compare to the probesets of repeat-free target sequences. In addition, 407 Affymetrix target sequences were incorrectly oriented relative to the genes they purportedly represented (anti-sense transcripts). Surprisingly, a large fraction of these "erroneous" sequences can be significantly associated with important regulatory biological processes, molecular functions and pathways. The all defined categories of probe sequences have been annotated in our local Affy Probes Mapping and Annotation (APMA) database. Our results allow us to re-identify many targets used in a microarray experiment and carry out biological classification of the anti-sense transcripts.

INTRODUCTION

Affymetrix GeneChip technology provides *in situ* synthesized oligonucleotide arrays with known sequence produced on each spot. GeneChip array uses a set (called probeset) of 11–20 oligonucleotide probes, each 25 bases long, to represent a gene. The expression level for a gene transcript is a sum of hybridization signals from the entire probeset. The perfect match probe comes together with a mismatch probe designed to measure specific hybridization signals (Affymetrix, 2004, http://www.affymetrix.com/support/). However, inadequate probe design and incorrect gene annotation has a clear potential to generate downstream problems for correct interpretation of microarray experiments. Recent papers (Mecham *et al*., 2004; Harbig *et al*., 2005) have re-evaluated of Affymetrix array probes

quality using BLAST alignment of probe sequences to the "complete" human genome. In this work, we combine sequence analysis of Affymetrix U133 original (target) sequences with clinical and biological validation of different categories of the target sequences. This approach allows us to re-evaluate the quality of many hundreds of Affymetrix target sequences and to obtain new knowledge on gene expression.

METHODS

Affymetrix sequence data for the U133A and U133B chips were downloaded from the NetAffx web site. These sequences, intended to represent genes, are referred to as "targets" of the Affymetrix probesets. We used these targets for the initial survey of possible transcripts that each probeset might detect. For validating accuracy of target sequences assignment, we used BLAT program, UCSC Genome Browser tools, and our additional programs developed at GIS and at BII. BLAT uses 90 %-complementarity criterion to match the target sequence for any genic region(s) of RefSeq, mRNA and splice variants on the HG17 assembly. These results were stored in a local database (APMA data base) associated with the probeset ID number from the chip. We also carried out manual curation and annotation of more than 2500 probesets representing the target sequences which were selected and classified by our in-house programs as problematic sequences. The Affymetrix target sequence is considered as a problematic sequence if it: (1) does not align by BLAT at 90 % complementary criterion in the human genome; (2) shows more than one match on the human genome; (3) shows an opposite orientation to genic sequence (anti-sense transcript), or to mRNA or most of ESTs in the EST cluster corresponding to the intended target.

In addition to this basic alignment and verification, we evaluated other potential complicating factors. For each sequence target we carried out a search for repetitive elements (using RepeatMasker) constructing table of repeats found by family and repeat types (DNA, LTR, LINE, SINE, simple and low complexity repeats, etc.). We calculated repeat coverage by percentage of the Affymetrix target sequence length. Split or chimerical probesets also were identified and flagged.

To validate usefulness of the problematic probes, we used statistical analysis of U133 Affymetrix microarrays on breast cancer tissue samples obtained from 260 primary breast cancer patients and stored in the database at GIS (Miller *et al*., 2005).

We used SAM (Statistical Analysis of Microarrays) software (Tusher *et al*., 2001) to estimate significance of differences in probesets expression level in biologically and clinically different groups of tumors (e.g. histological grade 1 and grade 3 breast cancers). For each of the Affy probesets "false positive rate" (q-value) was calculated by SAM program. We used Panther (http://www.pantherdb.org/) and DAVID (Dennis et al., 2003) Gene Ontology (GO) statistical software to estimate a significance of enrichment of specific gene categories in the groups of probesets studied.

RESULTS AND DISCUSSION

We revised the localization of Affy U133 sequences on the human genome. We found that: (1) 187 (0.42 %) probesets don't match any location in the human genome (internally called Tag0, see Table 1); (2) 42134 (94.3 %) probesets have unique genome location (reliable probesets, Tag1); (3) 2371 probesets (5.3 %) have multiple locations in the human genome, up to 10 times and more (Tag2, ...Tag11, etc.) and might cause potential cross-hybridization.

We believe that mismatched sequences (Tag0) and the sequences with multiple genome hits (called Tag2, Tag3, ... Tag11) are a source of uncertainties and cross-

hybridization affects in gene identification and should be excluded from analysis of array experiments.

Table 1.

# locations	Tag0	Tag1	Tag2	Tag3	Tag4	Tag5	Tag6	Tag7	Tag8	Tag9	Tag10	>= 11
#probesets	187	42134	1774	274	111	73	27	25	23	15	7	42
%	0.42	94.28	3.97	0.61	0.25	0.16	0.06	0.06	0.05	0.03	0.016	0.094

A novel and important feature of our analysis is a study of the repeat coverage of the transcripts represented by the target sequence. About 25 % of target sequences in Tag1 set are covered by mobile elements (repeats) abundant in the genome; hence, they might be a source of erroneous detection of expressed genes. Negative effect of repeats on gene expression level could be shown statistically on a large representative set of Affymetrix microarrays. We used an industry-standard GeneChip dataset of human breast cancer samples (Miller *et al*., 2005; database ID: NCBI GEO GSE3494).

Using a standard program to estimate of false positive rate, we selected a large number (~4000) of differentially expressed genes. We assume that if a given type of repeat elements covers a given target sequence, then corresponding probesets should be under-represented in a set of discriminating genes. We used a discrimination score which was estimated by a ratio of the numbers of differentially expressed genes (probesets) in the repeat-overlapped and the repeat-free sets. First, using SAM 2.1 program, we selected a list of differentially expressed probesets which can discriminate the low- and high-aggressive breast cancers at low "false positive rate" (q-value) of errors equals 1.5 %. Then, we counted the number of probesets with target sequence covered by a given type of repeats at 10 %, 20 %, . . . , 100 %. We found that as a general rule, and as expected, target sequences with repeats have progressively worsening significance for longer repeats (LTR and LINE) and for larger overlapping. Such proportion presents quality of probesets (Fig. 1).

Importantly, a quality level of an individual target sequences covered by repeats was typically reproduced across the clinical cohorts representing the patients from different hospitals and different countries (data not presented).

Additionally, we found that 407 Affymetrix probesets were designed using wrongly oriented sequences regarding to intended target gene (anti-sense transcripts). These probesets may have been designed based on poorly defined RNA sequences in which orientation was not defined accurately (e.g. EST clusters, pseudogene transcripts) and gene name had been assigned later; however, some may have originated from reverse-oriented artifact singleton cDNA clones whose incorrect orientation is evident when their directional genome alignments are compared to those of newer and more accurate cDNA sequences mapping to the same locus. Large fraction of ESTs/transcripts assigned to these sequences show low expression level.

Surprisingly however, many of "the wrongly oriented" probesets may not be useless. First, based on gene ontology (GO) Panther software tool, these sequences exhibited statistically significant enrichment by specific biological categories, relevant to cancer and signal transduction (*Biological process*: protein phosphorylation, protein modification, signal transduction, NF-kappaB cascade, cholesterol metabolism, MAPKKK cascade, oncogenesis, protein metabolism and modification. *Molecular Function*: Protein kinase, guanyl-nucleotide exchange factor, transcription cofactor, non-receptor serine/threonine protein kinase, G-protein modulator, protein kinase receptor, select regulatory molecule, tyrosine protein kinase receptor, nuclease).

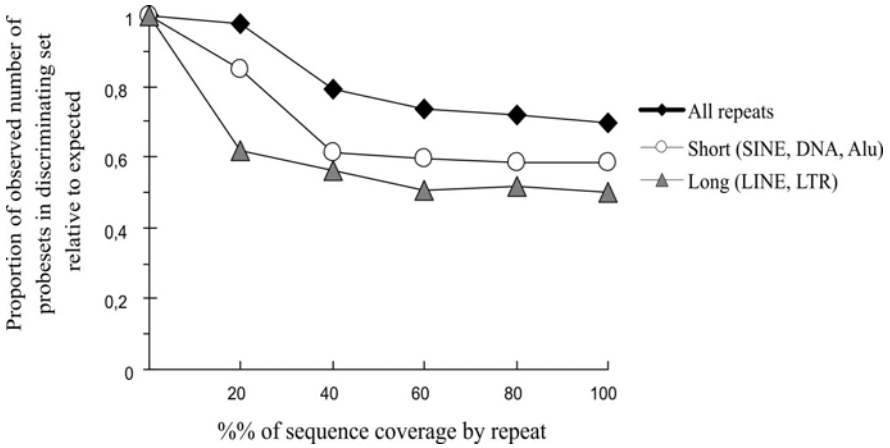


Figure 1. Estimations of the probability of occurrence of discriminating probesets (at q < 1.5 %) as a function of percent of target sequence span covered by repeat.

Second, using SAM program, we found that in comparison of 70 breast cancer samples classified as histological grade I tumor (low aggressive sub-type of breast cancer) versus 55 breast cancer samples classified as histological grade III tumor (high-aggressive sub-type of breast cancer), 58 probesets exhibit high confidence. 28 probesets were up-regulated in grade III tumors, and 30 probesets were up-regulated in grade I tumors (SAM q-value < 0.1). Statistical GO analysis of that confidence set of differentially expressed genes reveals enrichments of several biological categories associated with cancer. (Protein amino acid phosphorylation, cell cycle, mitosis, m-phases, nucleotide binding, protein kinase activity). These results suggest that a significant fraction of the wrongly oriented probesets detect real and tightly regulated transcripts. However, biological role of such transcripts from opposite strand at the same locus of a known gene has not been studied systematically.

Our results demonstrate that integration of statistical analysis of clinical data and genome-scale computational search of specific and reliable target sequences allows us to increase discovery potential of microarray data.

REFERENCES

Dennis G.Jr., Sherman B.T., Hosack D.A. *et al*. (2003) DAVID: Database for annotation, visualization, and integrated discovery. *Genome Biol*., **4**(5), 3.

Harbig J., Sprinkle R., Enkemann S.A. (2005) A sequence-based identification of the genes detected by probesets on the Affymetrix U133 plus 2.0 array. *Nucl. Acids Res*., **33**(3), e31.

Mecham B.H., Wetmore D.Z., Szallasi Z. *et al*. (2004) Increased measurement accuracy for sequence-verified microarray probes. *Physiol Genomics*, **18**(3), 308–15.

Miller L.D., Smeds J., George J. *et al*. (2005) An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proc. Natl. Acad. Sci. USA*, **102**(38), 13550–13555.

Tusher V.G., Tibshirani R., Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.

THE CONTENT OF miRNAS IN *ARABIDOPSIS THALIANA* CORRELATES WITH THE OCCURRENCE OF TETRAMERS WRHW AND DRYD

Ponomarenko M.P.**, Omelianchuk N.A., Katokhin A.V., Kolchanov N.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: pon@bionet.nsc.ru

Key words: miRNA, microRNA, nucleotide context, abundance

SUMMARY

Motivation: MicroRNAs (miRNAs) are short recently discovered non-protein-coding RNAs, which regulate gene expression.

Results: Using the system ACTIVITY to study the microarray data on the content of mature miRNA in *A. thaliana*, we found that a high content of miRNA correlates with a high occurrence of the tetranucleotides WRHW and DRYD in this miRNA sequence. It is shown that the linear regression of the unknown quantitative content of arbitrary miRNA on the basis of the known occurrences of the WRHW and DRYD within its sequence gives statistically significant predictions with independent control data. Till now functionally important context feature of mature miRNAs are unknown and here we first report that the sequence of mature miRNAs may also influence their ability to accumulate in tissues.

INTRODUCTION

MiRNAs are endogenous RNA with a length of 20–24 bases that bind in a complementary manner to messenger RNA (mRNA), which results in translation inhibition or destruction of the mRNA (Bartel *et al*., 2004). MiRNAs have different abundance and tissue specificity. A high content of miRNAs may be determined by both a high transcription level of their genes and a low rate of miRNA degradation. The findings in the present study suggest that the miRNA content in *A. thaliana* organs correlates with the occurrences of WRHW and DRYD tetranucleotides in the miRNA sequence.

METHODS AND ALGORITHMS

In the miRNA fragments (Table 1) with a length of 20 nt from the 5′-end ($\mathbf{E} = \mathbf{e_1...e_i...e_{L=20}}$, where nucleotide $\mathbf{e_i} \in \{A, U, G, C\}$) we study using the system ACTIVITY (Ponomarenko *et al*., 1999) the occurrence of oligonucleotides $\mathbf{Z(m)} = \mathbf{z_1...z_m}$ with a fixed length **m** ranging from 1 to 4 nt and weighted it taking into account their localization with the start at position **i** of this sequence:

$$X_{Z(m),F}(E)=\Sigma_{i=1,L-m+1}\, F(i)\, \Pi_{j=1,m}\, \Delta(e_{i+j-1} \in z_j), \quad (1)$$

where, $\mathbf{z_j} \in$ {A, U, G, C, W = {A,U}, R = {A,G}, M = {A,C}, K = {U,G}, Y = {U,C}, S = {G,C}, B = {U,G,C}, V = {A,G,C}, H = {A,U,C}, D = {A,U,G}, N={A,U,G,C}}; Δ(true) = 1, Δ(false) = 0; and **F(i)** is the weight function modeling the effect of **Z(m)** with a start at position **i** of the sequence **E** on the miRNA content using the heuristic rule "the higher is **F(i)**, the stronger is the influence of **Z(m)** at position **i** on the miRNA content. Overall, we analyzed 360 weight functions **F(i)** of two types (Fig. 1): 180 U-shaped **F(i)** with one peak (maximum or minimum) and 180 S-shaped **F(i)** with one transition (increase or decrease). Since till now, these functionally important context feature of mature miRNAs are unknown we made attempt to find these characteristics within miRNA sequences using U-shaped functions and gradients of these specific characteristics along miRNA sequences using S-shaped functions.

Table 1. Content of miRNAs in *A. thaliana*

miRNA	I	II	III	IV	V	VI	VII	VIII	WRHW	DRYD	Eq.(5)
mir158	3.590	3.889	**4.016**	**5.722**	**4.288**	5.851	**6.612**	**4.853**	1.790	2.013	5.780
mir159	**4.657**	5.237	4.179	5.788	5.076	6.003	5.528	**5.210**	1.300	1.399	4.329
mir160	3.469	**3.252**	2.085	**4.146**	3.363	4.678	5.699	**3.813**	0.685	2.434	4.881
mir161.1	4.373	4.331	3.655	4.756	4.703	**5.133**	**5.795**	4.678	**1.224**	**1.873**	4.852
mir161.2	2.999	3.517	**2.427**	4.531	**3.794**	4.392	**5.637**	3.900	**1.779**	**1.634**	5.267
mir163	**0.722**	**1.687**	0.739	**4.642**	**2.560**	1.308	2.049	**1.958**	0.608	1.700	3.815
mir164	**3.668**	**4.166**	**4.270**	**4.520**	4.187	**4.366**	**4.467**	4.235	**1.644**	**1.460**	4.861
mir165	0.658	0.722	**0.622**	**1.126**	0.717	**0.998**	**1.441**	**0.898**	0.000	0.700	1.702
mir166	1.494	**1.353**	**1.224**	**1.728**	**1.380**	1.522	**1.668**	**1.481**	0.000	0.700	1.702
mir167	4.929	2.132	**5.663**	5.639	**6.275**	**5.921**	4.134	4.956	2.226	1.172	**5.248**
mir168	**3.479**	**3.532**	**3.158**	4.349	**4.985**	**4.508**	**4.151**	**4.023**	1.000	1.249	3.738
mir170	**1.458**	**1.139**	1.153	**1.551**	1.652	**1.926**	-0.275	1.229	0.476	1.170	**2.945**
mir172	**6.257**	**7.769**	5.019	**7.774**	6.215	**3.966**	6.060	6.151	2.698	2.598	**7.742**
mir173	**0.786**	0.795	**-0.232**	0.945	**1.086**	**1.320**	0.830	0.790	1.610	0.549	**3.619**
mir390	**3.452**	**2.473**	**0.982**	2.888	3.557	2.718	**2.489**	**2.651**	0.434	2.128	4.149
mir394	**2.186**	3.014	1.505	2.174	1.204	2.534	1.400	2.003	0.000	0.247	**1.107**
mir396	**3.184**	**4.805**	2.856	**5.048**	**5.753**	4.266	4.167	**4.297**	1.262	0.549	3.162
mir156	1.182		**2.343**	3.230	**1.959**	**6.517**	**3.277**	3.085	**0.974**	**2.081**	4.797
mir169	**0.108**	1.548		**2.272**	2.367	**3.206**	3.159	2.110	0.000	1.611	**2.899**
mir171	3.215	**0.780**	2.379	1.551	**1.594**	**2.580**		**2.017**	0.476	1.170	2.945
mir398	0.115	**0.640**	**1.710**	**3.201**	**3.522**		**1.228**	1.736	**0.568**	**0.347**	1.985
mir156/157			–0.094	1.353	1.569	3.767	**0.299**		1.890	1.021	4.608
mir162	**0.116**			0.644	0.055				1.485	1.205	4.317
mir391				1.517	0.832				0.922	1.910	4.504
mir319	0.754	1.254							1.436	1.311	4.392
mir397b		-0.543		**0.940**	**0.216**		1.509		1.000	1.249	3.738
R	0.624	0.637	0.590	0.626	0.516	0.628	0.686	0.798	Fig.2*a*	Fig.2*b*	0.834
α	0.0025	0.0025	0.005	0.001	0.01	0.0025	0.0005	0.000025			0.05

Notes: Columns I–VIII: logarithmic units averaging according to Axtel and Bartel, 2005 and these data analysis provided by our approach. I, inflorescence; II, stem; III, silique; IV, cauline leaf; V, rosette leaf; VI, seedling; VII, root; VIII, averaged; R and α – linear correlation coefficient and its significance.

It is commonly known that statistical validation of the inference "the occurrence rate of $\mathbf{X_{Z(m),F}(E)}$ correlates with the content of [**miRNA**]" requires that all the pairs of variables $\{\mathbf{X_{Z(m),F}(E_k)}, \mathbf{[miRNA]_k}\}$ meet the condition of below simple regression:

$$[\text{miRNA}]_{Z(m),F}(E_k) = a + b * X_{Z(m),F}(E_k), \quad (2)$$

where, **a** and **b** are simple regression coefficients calculated conventionally according to the tested set of pairs of real numbers $\{\mathbf{X_{Z(m),F}(E_k)}, \mathbf{[miRNA]_k}\}$.

Eq. (2) utilizes the miRNA sequence $\mathbf{E_k}$ to predict the quantitative values $\mathbf{[miRNA]_{Z(m),F}(E_k)}$ indicating the content of this miRNA in *A. thaliana* based on the occurrence of **Z(m)** oligonucleotides in this miRNA. Regression (2) is applicable provided that there is a significant correlation between the predicted and experimental values—$\mathbf{[miRNA_{Z(m),F}(E_k)}$ and $\mathbf{[miRNA]_k}$. To test this, ACTIVITY first forms seven subsets of such pairs from all the analyzed pairs $\{\mathbf{X_{Z(m),F}(E_k)}, \mathbf{[miRNA]_k}\}$. Then for each of these 7 subsets, ACTIVITY tests 11 correlations between the prediction and the experiment. In particular, these 11 correlations include linear, sign, and rank correlations. Thus, ACTIVITY tests overall 7*11 = 77 partial correlations between the predicted and experimental values—$\mathbf{[miRNA_{Z(m),F}(E_k)}$ and $\mathbf{[miRNA]_k}$. Testing of each **n**th partial correlation ($1 \le \mathbf{n} \le 77$) consists in estimation of its significance α_n, which then is converted in terms of Zadeh's fuzzy logic (Zadeh, 1965) into validity estimate for the prediction tested:

$$q_n(X_{Z(m),F}(E) \rightarrow [miRNA]) = \begin{cases} 1, & \text{if } \alpha_n \le 0.01; \\ 1.3 - 28.3\alpha_n + 55.6\alpha_n^2, & \text{if } 0.1 \ge \alpha_n \ge 0.01; \\ -1, & \text{if } \alpha_n \ge 0.1. \end{cases} \quad (3)$$

Eq. (3) assigns to each significant correlation ($\alpha_n < 0.05$) between the predicted and experimental values, a positive validity estimate $\mathbf{q_n((X_{Z(m),F}(E) \rightarrow [miRNA])}$ ranging from 0 to 1; to each insignificant, the negative estimate ranging from –1 to 0. For each prediction $\{\mathbf{[miRNA]_{Z(m),F}(E_k)}\}$, it gives overall 77 partial validity estimates $\mathbf{q_n(X_{Z(m),F}(E) \rightarrow [miRNA])}$, which it averages into the integral validity estimate following the Decision Making Theory (Fishburn, 1970):

$$Q(X_{Z(m),F}(E) \rightarrow [miRNA]) = \{ \Sigma_{n=1,77}\, q_n(X_{Z(m),F}(E) \rightarrow [miRNA]) \}/77. \quad (4)$$

According to Eq. (4), the highest positive validity estimate $\mathbf{Q(X_{Z(m),F}(E) \rightarrow [miRNA])}$ indicates the particular oligonucleotide **Z(m)** and the particular weight function **F(i)** to predict (Eqs. 1 and 2) from the known miRNA sequences $\mathbf{\{E_k\}}$ the of miRNA content $\mathbf{[miRNA]_{Z(m),F}(E_k)}$ in *A. thaliana* that displayed the best fit with the experimental data.

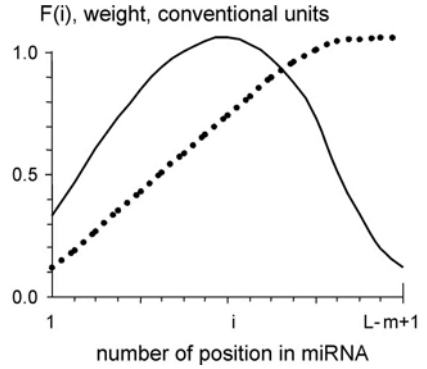


Figure 1. Examples of the weight functions **F(i)** used in Eq. (1). The most important **Z(m)** with the length m are at **i**th position of the central (solid line) and 3′-terminal (dotted line) parts of miRNA with the length **L**. Overall, 180 F(i) functions of that two types were used. That 360 weight functions with all the possible **Z(m)'s** with a length of $1 \le \mathbf{m} \le 4$nt allowed us to study $360*\{14+14*14+14*15*14+14*15*15*14\} \approx 2*10^7$ different quantitative variables $\mathbf{X_{Z(m),F}}$ calculated by Eq. (1) for any miRNA sequence.

RESULTS AND DISCUSSION

Using Eqs. (1)–(4), we analyzed the so-called training data subsets, amounting to 50 % of the overall data (Table 1, columns I–VIII, bold-faced), each representing uniformly all the experimental data studied. The rest 50 % of these data (regular font) were used as a control. For each analyzed miRNA with the sequence $\mathbf{E_k}$, $2*10^7$ weighted estimates of occurrences $\mathbf{X_{Z(m),F}(E_k)}$ were calculated by Eq. (1). For each $\mathbf{X_{Z(m),F}(E_k)}$, regressions (2) were constructed with the experimental data $\{\mathbf{[miRNA]_k}\}$ (Table 1, columns I–VIII, bold-faced) to predict the miRNA content $\{\mathbf{[miRNA]_{Z(m),F}(E_k)}\}$ by their sequences $\mathbf{E_k}$. Then Eqs. (3) and (4) allowed for deriving the validity estimates $\mathbf{Q(X_{Z(m),F}(E)\rightarrow[miRNA])}$ for each of these predictions. Overall, 10^8 of such estimates $\mathbf{Q(X_{Z(m),F}(E)\rightarrow[miRNA])}$ were obtained for eight training subsets (table 1, columns I–VIII, bold-faced). In the case of U-shaped weight functions, the regression (2) of the average miRNA content (Table 1, column VIII) according to the occurrence $\mathbf{X_{WRHW,F1}(E)}$ of WRHW weighted with the function $\mathbf{F_1(i)}$, having its maximum in the center of miRNA (Fig. 1, continuous line), was the most valid, $\mathbf{Q}$=0.477. The column "WRHW" lists the values $\mathbf{X_{WRHW,F1}(E_k)}$ calculated using Eq. (1) for all the miRNA studied.

The correlation between these $\mathbf{X_{WRHW,F1}(E_k)}$ and the average miRNA content $\{\mathbf{[miRNA]_k}\}_{VIII}$ is shown in Fig. 2*a*. For S-shaped weight functions, the regression (2) of the miRNA content in roots (Table 1, column VII) according to the occurrence rate $\mathbf{X_{DRYD,F2}(E)}$ of DRYD weighted with the function $\mathbf{F_2(i)}$ given in Fig. 1 (dotted line) was the most valid, $\mathbf{Q}$ = 0.466. The column "DRYD" lists all the values $\mathbf{X_{DRYD,F2}(E_k)}$; Fig. 2*b* shows the correlation between $\mathbf{X_{DRYD,F2}(E)}$ and miRNA content in *A. thaliana* roots, $\{\mathbf{[miRNA]_k}\}_{VII}$.

Since the WRHW and DRYD occurrences were independent ($\mathbf{R}$ = 0.324, $\alpha > 0.10$), the linear regression of the average miRNA content in *A. thaliana* by has been standard made:

$$[\text{miRNA}](E_k) = 0.782 + 1.314*X_{WRHW,F1}(E_k) + 0.756*X_{DRYD,F2}(E_k). \quad (5)$$

To this end, the pairs $\{\mathbf{X_{WRHW,F1}(E_k)}, \mathbf{X_{DRYD,F2}(E_k)}\}$ bold-faced in the Table 1, columns "WRHW" and "DRYD" and the corresponding experimental values $\{\mathbf{[miRNA]_k}\}_{VIII}$ (column VIII) were used. The column "Eq.(5)" lists the all values $\mathbf{[miRNA](E)}$ predicted. Bold-faced in this column are predictions for six miRNA that were not previously used for either optimizing Eq. (5) or search for the WRHW and DRYD (columns "WRHW", "DRYD", VII, and VIII, regular font). The bottom line of column "Prediction" contains the linear correlation coefficient $\mathbf{R}$ = 0.834 between these six independent predictions and the corresponding experimental values of miRNA average content in *A. thaliana*. These control predictions fit significantly the experimental data (α<0.05). In addition, the two bottom lines in Table 1 show the linear correlation coefficients between the predictions according to Eq. (5) and experimentally determined miRNA contents in all the seven *A. thaliana* organs studied. As is evident, these all correlations are statistically significant ($\alpha < 0.01$).

The difference in nucleotide context in 3′-end of miRNAs may attribute to the different binding of this part of miRNA sequences to the PAZ domain of an Argonaute protein, the core constituent of the RISC (Tomari *et al*., 2004) and nucleotide context of the central part in this case provides a different flexibility between the tightly bound 3′ end and the 5' half of the small RNA pre-organized for binding an RNA target (Tomari, Zamore, 2005).

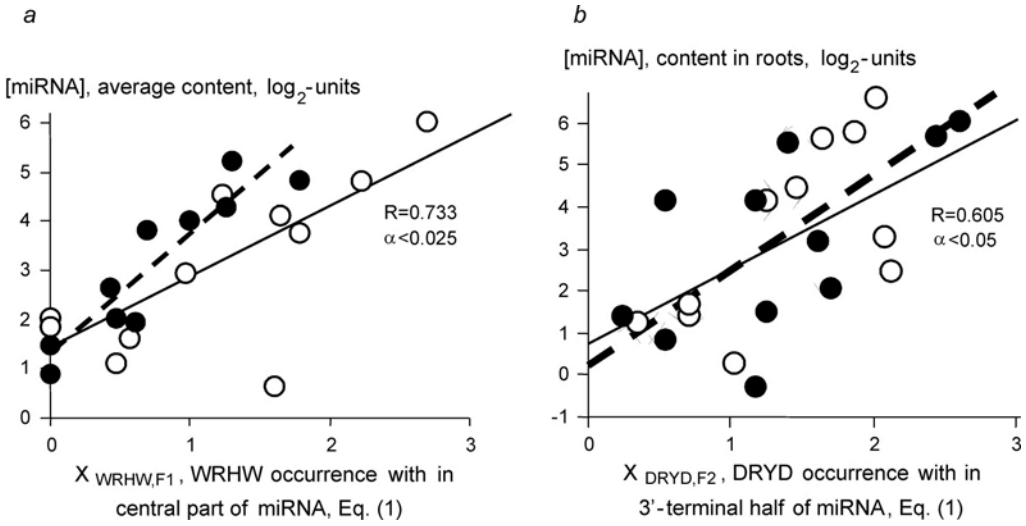


Figure 2. Contextual patterns of miRNA. *a* – the average content of miRNA in *A. thaliana* (vertical axis, experiment) correlates with the occurrence of WRHW weighted with the function F_1 (Fig. 1, continuous line) in miRNA sequence (horizontal axis, prediction). *b* – the miRNA content in *A. thaliana* roots (vertical axis, experiment) correlates with the occurrence of DRYD weighted with the function F_2 (Fig. 1, dotted line) in miRNA sequence (horizontal axis, prediction). Notes: dark circles and dotted line for training data and light circles and continuous line, for control data; **R** and **α** – linear correlation coefficient and its significance.

ACKNOWLEDGEMENTS

This work was supported by Russian Federal Agency of Science and innovation (IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)" and Living system program award "Identification of potential targets for novel medicinal drugs based on reconstructed gene networks").

REFERENCES

Axtell M.J., Bartel D.P. (2005) Antiquity of microRNAs and their targets in land plants. *Plant Cell*, **17**, 1658–1673.

Bartel D. (2004) MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Fishburn P.C. (1970) Utility theory for decision making. New York: Jonh Wiley & Sons.

Ponomarenko M. *et al*. (1999) Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins. *Bioinformatics*, **15**, 687–703.

Tomari Y. *et al*. (2004) A protein sensor for siRNA asymmetry. *Science*, **306**, 1377–1380.

Tomari Y., Zamore P.D. (2005) Perspective: machines for RNAi. *Genes Dev*., **19**, 517–529.

Zadeh L.A. (1965) Fuzzi sets. *Information and Control*, **8**, 338–353.

THE ANALYSIS OF SREBP BINDING SITES DISTRIBUTION IN GENE REGIONS BY COMBINED SITEGA AND PWM APPROACH

***Proskura A.L.*[*1], *Levitsky V.G.*[1, 2], *Ignatieva E.V.*[1, 2]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: anya@bionet.nsc.ru

Key words: lipid metabolism, position weight matrix, discriminant analysis, genetic algorithm

SUMMARY

Motivation: Disruption of lipid metabolism is known to cause a set of severe human diseases. Transcription factor SREBP (Sterol Regulatory Element Binding Protein) is a key regulator of cholesterol homeostasis gene expression, hence, analysis of SREBP binding site data, as well as development of reliable methods for SREBP recognition are extremely important tasks.

Results: For SREBP recognition we applied a combined approach. The approach compiled the SiteGA method that was implemented using a genetic algorithm involving a discriminant analysis of dinucleotide context and position weight matrix (PWM) method. We have studied distribution of potential SREBP binding sites (BS) in promoters, exons and introns of lipid-specific genes (LM-TRRD) and the EPD based set of human promoters. The highest densities of predicted SREBP BS were observed for promoters and introns of lipid-specific genes. The combined approach application may overcome the drawbacks of individual methods thereby the most reliable SREBP targets may be found.

INTRODUCTION

Transcription factors of the SREBP family play an important role in regulation of expression of genes controlling cholesterol level and synthesis of triglycerides in a cell, hence, analysis of SREBP binding site data, as well as development of a reliable methods for SREBP binding sites recognition are extremely important tasks. The active SREBP form is obtained from inactive precursor, this process being suppressed by increasing inner cellular cholesterol level (Brown, Goldstein, 1997). As known, the factors of this family, SREBP1a, SREBP1c, and SREBP2, belong to the family of bHLHLZ (basic helix-loop-helix leucine zipper) proteins and bind to the sites of E-box and SRE (non-E-box) types. By taking into account the differences in the context organization of SREBP sites, it seemed reasonable to develop recognition methods for each of two sub-types individually. Moreover, the recognition accuracy for BS of E-box type appeared to be too low. The SREBP BS recognition is hampered by high false positive rate (Proskura *et al.*, 2004), that's why it is very difficult to get success in SREBP BS large-scale genome research. In order to decrease false positive rate we combined SiteGA (Levitsky *et al.*, 2006) and PWM (Stormo *et al.*, 2000) site recognition methods in our analysis. We observed differences in BSs density for lipid-specific and other non-specific nucleotide sequences.

METHODS AND ALGORITHMS

In our analysis we used SREBP BS of SRE-type. The nucleotide sequence sets used in our analysis presented in Table 1. The BSs sequences with flanks with centrally located BS were compiled in SREBP train set. This set was extracted from the lipid metabolism section LM-TRRD of the TRRD database (Kolchanov *et al.*, 2002). The promoters of genes of the lipid metabolism system were also compiled from the LM-TRRD database. All sequences contained sites from train set were removed from LM-TRRD set. The SREBP control set was derived from literature sources and it was used for thresholds setting. For both methods these settings corresponded to 50 % of true positive rates, which were estimated by the control set. LM exons and LM introns sets were extracted directly from EMBL on the basis of information stored in LM-TRRD. In our analysis we also used the human promoters gene regions that were extracted from EPD database (Schmid *et al.*, 2006). The random sequence set was obtained by shuffling of the train SREBP set (i.e. nucleotide content remained the same).

Table 1. Samples of nucleotide sequences

Sample name	Sequence length, nt	No. of sequence
SREBP train	18	38
SREBP control	18	8
LM-TRRD promoters, [-1000;+100][1]	1100	82
LM exons	56182[2]	292
LM introns	172778[2]	209
EPD promoters, [-1000; +100][1]	600	1871
Random sequences	900	1900

[1] location relative to transcription start site, lacking the 5′- or 3′-flanks of nucleotide sequences completed with the symbol "n"; [2] total analyzed length (the number of 18-nt window locations for both DNA strands).

We used SiteGA (Levitsky *et al.*, 2006) and mononucleotide PWM (Stormo, 2000) recognition methods in our analysis. The description of both methods implementation may be found elsewhere (Levitsky *et al.*, this issue). First of all in our analysis we applied each method separately, than we combined them. That means that combined method considered analyzed sequence as a potential site if it was predicted simultaneously by both methods.

RESULTS AND DISCUSSION

Firstly we compared the recognition performance of SiteGA, PWM and combined SiteGA & PWM recognition methods. The control BS sequences were used for estimation of dependences of false positives (FP) vs. false negatives (FN) (Fig. 1).

In general, three methods had the similar recognition performance. Nevertheless, it may be suspected that combined method may outperform others in the low false positive rate area, which corresponds to prediction of the most reliable BS targets.

Than we considered SiteGA, PWM and combined SiteGA & PWM recognition methods application. We expected the highest SREBP BS densities in the LM-TRRD and introns sets. In contrast to these, the densities in the LM exons and random sets were suspected to be reduced. The EPD set may be considered as another independent negative background, that may contain less putative BSs than LM-TRRD set. Fig. 2 presents the results of these data analysis.

As expected combined approach gave the highest potential site densities for LM-TRRD and LM introns sets. On the contrary, as it was expected the lowest densities were observed for LM exons and EPD sets. The different densities of predicted by SiteGA and

PWM methods sites were found for random set. Moreover, application of PWM method for random set gave the highest density among all other sets tested by this method. This may be explained as a common trend of PWM methods to be sensitive to the nucleotide content of analysed sequences. The differences between two methods found for other nucleotide sequences sets were not so obvious (Fig. 2).

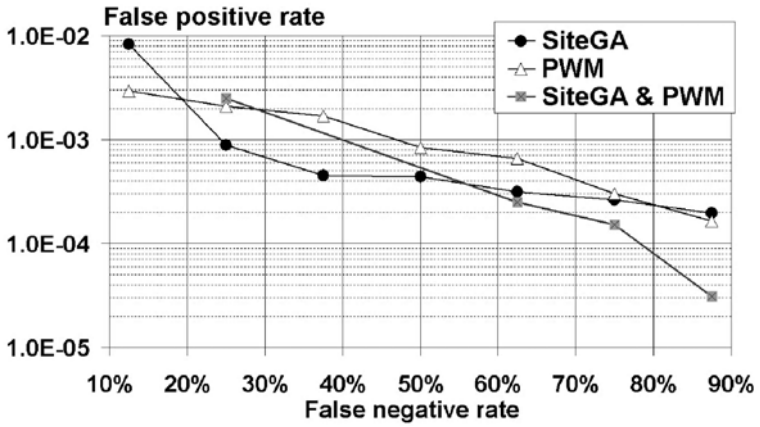


Figure 1. Comparison of the recognition performance of the SiteGA, PWM combined SiteGA & PWM methods, estimated by control data set.

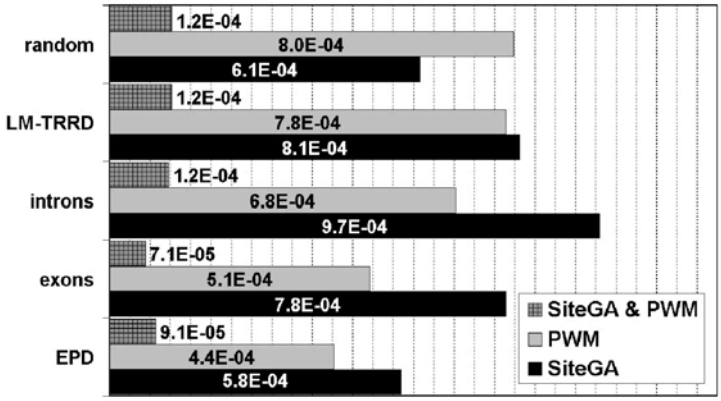


Figure 2. The distribution of potential SREBP BS densities in regulatory regions, exons and introns of lipid-specific genes (LM-TRRD) and the EPD promoters.

Finally to ensure the predictive capabilities of combined SiteGA & PWM method against each one taken separately we calculates the ratios of predicted site densities at different stringencies for LM-TRRD and EPD sets (Fig. 3).

The LM-TRRD and EPD sets we considered correspondingly as ‘YES’ and “NO”, since the first one in comparison with the second contained more potential SREBP sites. To calculate ratios YES/NO we used the presented above control data (Table 1), i.e. thresholds used for SiteGA and PWM methods corresponded to consequently from 1 to 7 control BSs predictions. It may be concluded, that at least for two most stringent thresholds combined SiteGA & PWM method appeared to be significantly better than each method taken separately. This is also confirmed by the accuracy estimation (Fig. 1). Moreover, the

combined method may correct some drawbacks of separate methods (for example this refers to the tendency of PWM method to find comparatively many sites in random set, Fig. 2).

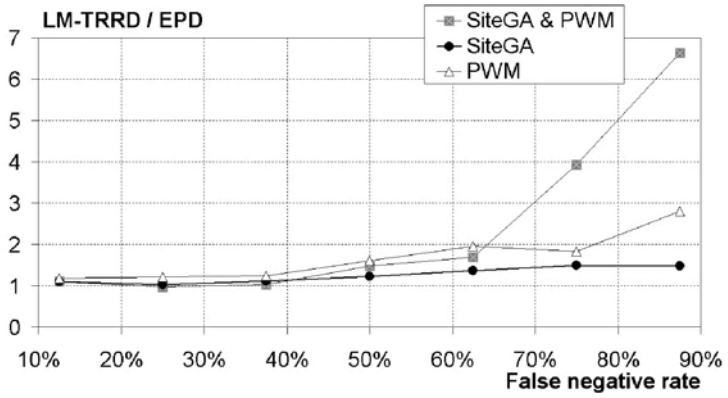


Figure 3. The ratios of potential SREBP BS densities in regulatory regions of lipid-specific genes (LM-TRRD) and the EPD genes.

Finally we may conclude that combined approach application may reveal most reliable SREBP targets. This conclusion may be considered as the strategy for large-scale genome research in the case when the high false positive rate don't allow to reach the appropriate recognition performance.

ACKNOWLEDGEMENTS

The work was supported by the RFBR (grant No. 05-04-49111); the Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 43.073.1.1.1501); Innovation Project of Federal Agency of Science and innovation IT-CP.5/001 "Development of software for computer modelling and design in postgenomic system biology (system biology *in silico*); U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF) and the Ministry of Education of Russian Federation within the Basic Research and Higher Education Program (Award No. REC-008, grant Y2-B-08-02). The authors are grateful to Lokhova I.V. for technical support.

REFERENCES

Brown M.S., Goldstein J.L. (1997) The SREBP Pathway: regulation of Cholesterol metabolism by proteolysis of a membrane-bound transcription factor. *Cell*, **89,** 331–340.

Kolchanov, N.A. *et al.* (2002) Transcription regulatory regions database, (TRRD): its status in 2002. *Nucl. Acid Res.*, **30**, 312–317.

Levitsky V.G. *et al.* (2006) The SiteGA and PWM methods application for transcription factor binding sites recognition in EPD promoters. *This issue*.

Levitsky V.G. *et al.* (2006) Method SiteGA for transcription factor binding sites recognition. *Biofizika* (in press).

Proskura A.L. *et al.* (2004) SREBP binding sites: context features and analysis of genome distribution by the SITECON method. *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2004),* **1**, 174–178.

Schmid C.D. *et al.* (2006) EPD in its twentieth year: towards complete promoter coverage of selected model organisms. *Nucl. Acids Res.*, **34**, 82–85.

Stormo G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

SIMPLE SEQUENCE $(TG/CA)_N$ REPEATS AS CIS MODULATORS OF GENE EXPRESSION

Ramachandran S.*, Sharma V.K., Sharma A., Brahmachari S.K.
G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi, 110 007, India
* Corresponding author: e-mail: ramuigib@gmail.com

Key words: gene expression, regulation, repeats, human genome

SUMMARY

Motivation: The $(TG/CA)_n$ repeats display polymorphic properties and exhibit *cis* regulatory characteristics. Analysis of distribution and regulatory effects of these repeats could provide insights into their role in the regulation of genome wide expression in the human genome.

Results: The ratio of the number of genes with $(TG/CA)_n$ repeats to the total number of genes is uniform across all human chromosomes. The number of genes with repeats decreased with increasing repeat length and several genes (53 %) had multiple types of repeats in various combinations. Signalling and communication genes were rich with repeats whereas the genes of Immune and related functions and Information were poor in repeats. Proportion of genes in a functional group with repeats bears a linear relation to gene length. Most repeats are located in introns. Incidence of repeats caused lowering of transcript levels. These results were observed in independent microarray datasets and corroborates with single gene studies.

Availability: http://expoldb.igib.res.in/expol.

INTRODUCTION

A potential regulator of transcription in eukaryotes is the dinucleotide $(TG/CA)_n$ repeat. The $(TG/CA)_n$ repeats are abundant in the human genome, polymorphic ($n \geq 12$) and act as *cis* regulators of transcription (Sharma *et al.*, 2003). In addition, these repeats have been observed to be associated with recombination sites (Majewski, Ott 2000) and mRNA splicing (Hui *et al.*, 2003) which elect them as functional elements (Sharma *et al.*, 2005). The $(TG/CA)_n$ repeats can be categorized into three types, Type I, Type II and Type III based on their length and biological properties. Type I repeats ($6 \leq n < 12$) are short, and have very low propensity for polymorphism. Type II repeats ($12 \leq n < 23$) are likely polymorphic, as more than 93 % of the $(CA)_n$ repeats of $n \geq 12$ units were found to display length polymorphism and act as *cis* regulators of transcription (Dib *et al.*, 1996; Sharma *et al.*, 2003). The Type III repeats ($n \leq 23$) were shown to have a propensity to adopt conformations such as Z DNA (Haniford, Pulleyblank, 1983; Nordheim, Rich, 1983; Peck, Wang, 1985; Meera *et al.*, 1989) and were shown to be associated with recombination sites (Majewski, Ott, 2000). In general, $(TG/CA)_n$ repeats of $n \geq 12$ units exert a down regulatory effect on transcription, which is positively correlated with the length of repeats (Agarwal *et al.*, 2000). A few examples of genes, whose transcription levels were shown to be modulated by $(TG/CA)_n$ repeats are summarized in (Sharma *et al.*, 2003; 2005). A model for the mechanistic role of $(TG/CA)_n$ repeats is shown in Fig. 1.

Recently genome wide expression technologies have generated data using microarrays. We present our studies using these data.

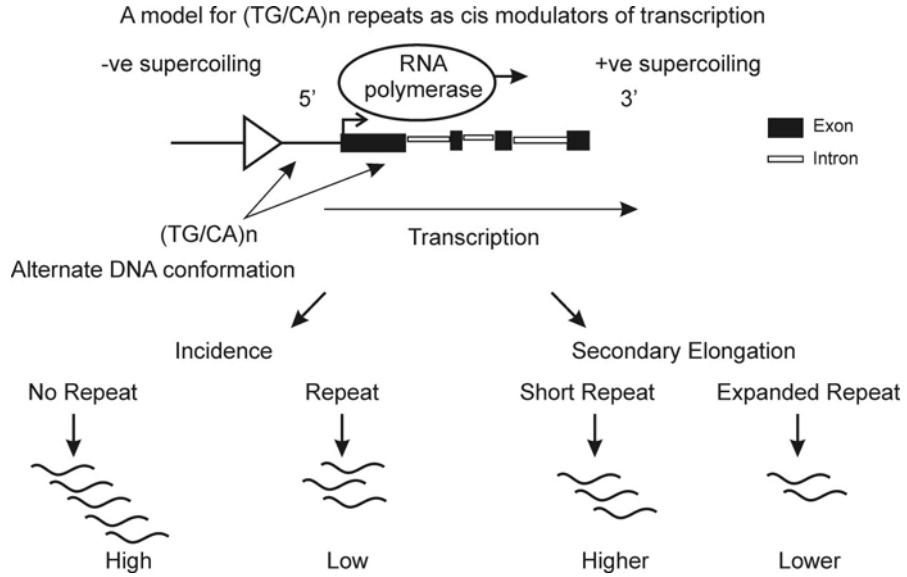


Figure 1. A model for the mechanistic role of (TG/CA)n repeats as cis modulators of transcription. The RNA polymerase complex generates positive supercoiling ahead of it and negative supercoiling behind as it ploughs along the template DNA during transcription. In this movement, if it encounters (TG/CA)n repeats with propensity to adopt conformations other than the usual B form, transcription is affected. In most cases studied so far, this effect is that of retardation, reducing the amount of transcripts generated. Two forms of polymorphism are observed: Incidence polymorphism involves accretion or degeneration of repeats whereas secondary elongation involves expansion or contraction of repeats.

METHODS AND ALGORITHMS

Human Gene sequences were retrieved from http://www.ncbi.nlm.nih.gov/LocusLink/. Uninterrupted Repeats that are more likely to be polymorphic were identified using the computer program SimRep (Sharma *et al.*, 2003). Clustering of genes into families was carried out using the root symbol assigned by the Hugo Gene Nomenclature Committee (Sharma *et al.*, 2005). Microarray datasets were collected from Hsiao *et al.* (2001) and Sharma *et al.* (2005b).

RESULTS AND DISCUSSION

The ratio of the number of genes with repeats to the total number of genes in all chromosomes falls in the narrow range 0.43 to 0.70 with a mean value of 0.59 indicating a near uniform distribution across all the chromosomes. The functional class of Signaling and communication had the highest number of genes with repeats and was significantly higher than the expected proportion assuming no preference for any of the functional classes ($P < 0.0001$). On the other hand, the classes of Immune and related information and Information had significantly lower proportion of genes with repeats ($P < 0.0001$). We also observed that the differences in the proportion of genes with repeats between the various functional classes is controlled more strongly by function than GC content, which varies in the narrow range 47–49 % whereas the proportion of genes with repeats varies widely in the range 29.6–61 %. Furthermore, there is a significant positive correlation

between the average gene length in each of the functional classes and the proportion of genes with repeats (R = 0.93, $P < 0.007$). The proportion of genes containing repeats decrease in the order type I > type II > type III. This trend was observed in all functional classes. These observations indicate that indefinite expansion of repeats is disallowed. The abundance of short repeats and rare occurrences of long repeats suggests a power law type relationship. It is well known that power law relation is now observed in many genomic properties and the patterns of repeat distribution also likely fit to power law trend. Short repeats (type I) are however not likely to be polymorphic and several of them could be on way to decay by accumulating mutations. Using microarray expression datasets, comparison of genes with similar expression patterns such as Housekeeping genes revealed that genes with repeats (type II and III) have lower transcriptional levels compared to those without repeats (n < 6 units) (t-test, $P < 0.0001$). These results corroborate the observations on the role of repeats in single gene studies. We are now in the process of preparing a database EXPOLDB dedicated to facilitate these investigations using either gene centric or pathway centric approach.

ACKNOWLEDGEMENTS

SR and SKB are recipients of grants "*In Silico* Biology for Drug target identification" (CMM0017) and New Millennium Indian Technology Leadership Initiative (NMITLI) from CSIR and HP centre for excellence. VKS and AS are recipients of fellowships from CSIR.

REFERENCES

Agarwal A.K., Giacchetti G., Lavery G., Nikkila H., Palermo M., Ricketts M., McTernan C., Bianchi G., Manunta P., Strazzullo P., Mantero F., White P.C., Stewart P.M. (2000) CA-Repeat polymorphism in intron 1 of *HSD11B2*: effects on gene expression and salt sensitivity. *Hypertension*, **36**, 187–194.

Dib C., Faure S., Fizames C., Samson D., Drouot N., Vignal A., Millasseau P., Marc S., Hazan J., Seboun E., Lathrop M., Gyapay G., Morissette J., Weissenbach J. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, 152–154.

Haniford D.B., Pulleyblank D.E. (1983) The *in vivo* occurrence of Z DNA. *J. Biomol. Struct. Dyn.*, **1,** 593–609.

Hui J., Stangl K., Lane W.S., Bindereif A. (2003) HnRNP L stimulates splicing of the eNOS gene by binding to variable-length CA repeats. *Nat. Struct. Biol.*, **10**, 33–37.

Hsiao L.L., Dangond F., Yoshida T., Hong R., Jensen R.V., Misra J., Dillon W., Lee K.F., Clark K.E., Haverty P., Weng Z., Mutter G.L., Frosch M.P., Macdonald M.E., Milford E.L., Crum C.P., Bueno R., Pratt R.E., Mahadevappa M., Warrington J.A., Stephanopoulos G., Stephanopoulos G., Gullans S.R. (2001) A compendium of gene expression in normal human tissues. *Physiol. Genomics*, **7**, 97–104.

Majewski J., Ott J. (2000) GT Repeats are associated with recombination on human chromosome 22. *Genome Res.*, **10**, 1108–1114.

Meera G., Ramesh N., Brahmachari S.K. (1989) Zintrons in rat alpha-lactalbumin gene. *FEBS Lett.*, **251**, 245–249.

Nordheim A., Rich A. (1983) The sequence (dC-dA)n X (dG-dT)n forms left-handed Z-DNA in negatively supercoiled plasmids. *Proc. Natl. Acad. Sci. USA*, **80**, 1821–1825.

Peck L.J., Wang J.C. (1985) Transcriptional block caused by a negative supercoiling induced structural change in an alternating CG sequence. *Cell*, **1**, 129–137.

Sharma V.K., B-Rao C., Sharma A., Brahmachari S.K., Ramachandran S. (2003) (TG:CA)(n) repeats in human housekeeping genes. *J. Biomol. Struct.Dyn.*, **21**, 303–310.

Sharma V.K., Brahmachari S.K., Ramachandran S. (2005a) (TG/CA)n repeats in human gene families: abundance and selective patterns of distribution according to function and gene length. *BMC Genomics*, **6**, 83.

Sharma A., Sharma V.K., Horn-Saban S., Lancet D., Ramachandran S., Brahmachari S.K. (2005b) Assessing natural variations in gene expression in humans by comparing with monozygotic twins using microarrays. *Physiol Genomics*, **21**, 117–123.

IDENTIFICATION OF MicroRNAS ENCODED BY *DROSOPHILA* TRASPOSABLE ELEMENTS

Ryazansky S.S.*
Institute of Molecular Genetics, RAS, Moscow, Russia
* Corresponding author: e-mail ryazansky@img.ras.ru

Key words: microRNA, transposable elements, *Drosophila*

SUMMARY

Motivation: microRNAs are small 21–24 nt. single stranded RNA molecules that capable to silence gene expression on a posttranscriptional level. MicroRNAs are involved in a regulation of a significant part of eukaryotic genes. Several bioinformatics approaches were explored to identify microRNAs encoded genes in genomes of various eukaryotes (including *Drosophila melanogaster*) and viruses (Bartel, 2003; Kong, Han, 2005). However, analysis of heterochromatin and interspersed repetitive elements has not been performed. Here we attempted to fill this gap and succeeded to find *in silico* microRNAs encoded by transposable elements (TEs) of *Drosophila*.

Results: Canonical sequences of *D. melanogaster* (Dm) TEs were tested using previously developed methods for microRNAs identification. As a result several microRNAs and their precursors were predicted. Now we are trying to detect them *in vivo*.

INTRODUCTION

All known bioinformatics approaches of microRNAs identification are based on two main microRNAs features (Kong, Han, 2005): 1) as a rule, microRNAs are conserved among eukaryotes; 2) microRNAs precursors (pre-miRNAs) have well-defined and stable secondary structure (60–90 nt. hairpin). It was also noted that pre-miRNAs have defined GC % composition; for example, pre-miRNAs of Dm are characterized by 31.9–59.3 % GC content (Bonnet *et al*., 2004). To identify the microRNAs encoded by TEs of Dm we scanned TEs sequences to find hairpin-like structures; after that, the obtained candidate pre-miRNA hairpins were filtered by GC content and hairpin stability; then, we searched possible homologues of candidate hairpins in Ds and Dy genomes and verified their ability to form hairpins; finally, potential microRNAs of selected hairpins were determined by MiRscan program.

METHODS AND ALGORITHMS

Canonical sequences of TEs were downloaded from BDGP server (http://www.fruitfly.org/p_disrupt/datasets/NATURAL_TRANSPOSABLE_ELEMENTS.fa). Potential pre-miRNA hairpins were searched by *srnaloop* program in both 5′ and 3′ strands (Grad *et al*., 2003) with the following parameters: `-w 4 -dw 1 -~2w -t 30 -l 110 -sw 0`. GC % of candidate hairpins was tested by written Perl script. Hairpin stability was verified by *randfold* (Bonnet *et al*., 2004) and sequences with $p < 0.01$ were collected. The search of hairpin homologues was done using WU-BLAST

(http://www.genome.wustl.edu/tools/blast) against Ds and Dy genomes; alignments were carried out by ClustalX (Thompson *et al*., 1997). Conservation of RNA hairpins were checked by *alifold* (Hofacker *et al*., 2002). Sequences of potentially microRNAs were searched by web-available version of MiRscan (http://genes.mit.edu/mirscan, Lim *et al*., 2003). *Drosophila melanogaster* GenBank EST sequences were analyzed by BLAST on the Flybase server (http://flybase.org/blast/).

RESULTS AND DISCUSSION

Using *srnaloop* and filters we identified twenty conserved regions (~100 bp.) possible to transcribe into RNA with stable hairpin structure. The quality of these hairpins determined by *alifold* is evaluated. MiRscan was developed as computational tool to identify specifically microRNAs in hairpins that are conserved in two genomes and have the features of known microRNAs. The results of MiRscan analysis are presented as score values; the majority of known microRNAs has a score more then 10 (Lim *et al*., 2003). Table 1 shows the TE regions forming conservative hairpins with a score more than 10 at least in one pair of homologues (Dm/Ds and/or Dm/Dy). This list represents 11 hairpins that appear to be pre-miRNAs. Interestingly, TEs encoding predicted pre-miRNAs belong only to the LTR and LINE classes, but not to TIR. Sequences from Cr1A, Ivk and R1-element have high score in both Dm/Ds and Dm/Dy pairs (see Fig. 1).

Table 1. Results of *alifold* and MiRscan analyses and searching in dbEST

TE	Hairpin position*, bp.	Class	*alifold*	MiRscan score		ESTs**
				Dm/Ds	Dm/Dy	
Cr1A	3′ 2492–2601	LINE	++	**16,9**	**15,4**	+,-
Ivk	3′ 2488–2586 (*pol*)	LINE	+++	**14,1**	**12,3**	none
R1-element	5′ 1061–1170 (CDS 1)	LINE	+	**13**	**11**	+,-
MAX-element	3′ 1723–1832 (border 5′UTR-ORF1)	LTR	+++	**15**	7	-
blood	3′ 5792–5901 (CDS 3)	LTR	+++	**12,4**	-4,5	+,-
Het-A	3′ 5125–5232 (3` UTR)	LINE	++	**10,1**	-15,6	+,-
springer	5′ 1563–1672 (CDS 2)	LTR	+++	**18,4**	-0,1	none
Quasimodo	3′ 6858–6965 (LTR)	LTR	++	**17**	-19	+,-
Cr1A	3′ 837–946	LINE	+++	2,8	**17,5**	+,-
BS	5′ 2815–2924, 3′ 6410–6519	LINE	++	6,3	**10,2**	none
gypsy6	5′ 1818–1927	LTR	+++	-1	**17**	none

Notes: * 5′ and 3′ indicate the corresponding strand of TE where *srnaloop* found the hairpin. ** + and – indicate the presence of hits in the same and in the opposite orientation as compared to the query respectively; none indicates the absence of hits.

Five hairpins from MAX-element, blood, HeT-A, springer and Quasimodo have score more than 10 only in Dm/Ds pair; the rest hairpins from Cr1A, BS and gypsy6 have high score only in Dm/Dy pair of possible homologues.

It is generally considered that evolution of TEs differs from evolution of the whole host genome due to their horizontal transfer between *Drosophila* species. Probably, this selective pattern of the proposed microRNA-genes conservation (conserved only in Dm/Ds or Dm/Dy pairs or both) is linked with phylogenetic peculiarities of the corresponding TEs in Ds, Dy and Dm species. This assumption now is checked.

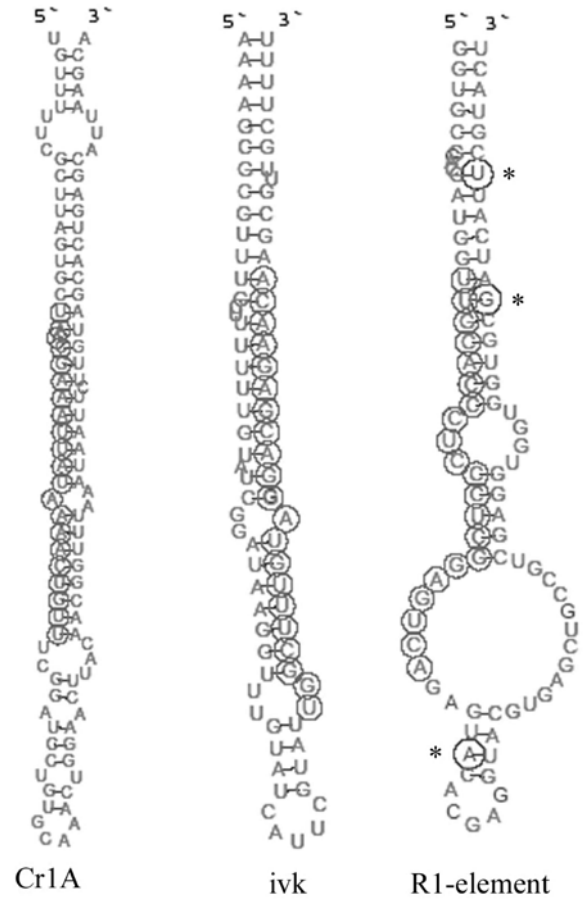


Figure 1. Candidate pre-miRNAs with the highest MiRscan score. The images were generated by MiRscan program for Dm/Ds pair. The hairpin residues of predicted microRNAs are circled. The circled residues with asterisk indicate the nucleotide positions where the sequence of R1-element in Dm and Ds are differing.

At least four proposed pre-microRNAs reside in CDS regions (Table 1): candidates from R1-element and springer are predicted to be on the sense strand; candidates from Ivk and blood – on the antisense strand. It is in agreement with early observation that microRNAs can be encoded within exons of mRNA encoded genes (Berezikov *et al*., 2005). The hairpins from MAX-element, HeT-A and Quasimodo reside in the non-coding regions. The annotation for the others TEs isn't available.

To support the expression of candidate pre-miRNAs we searched the corresponding ESTs in the *Dm* dbEST (Table 1). Six candidates have hits in the same orientation as the query sequence and one (MAX-element) have hits corresponding only to the opposite strand of the region forming the proposed hairpin; the others have not any hits. It can to speculate, that the antisense transcription of TEs can be initiated by some intrinsic antisense TE-promoters or nearest non-TE promoters. The antisense transcription of several TEs was shown *in vivo* by RT-PCR (Klenov and Ryazansky, ms. in preparation). Probably, the absence of hits in dbEST for the several candidates can be explained by their typical for the most of known miRNAs expression in the time- and tissue-specific manner and the underrepresented *Drosophila* dbEST.

To ensure the expression of the predicted pr-microRNAs and microRNAs we plan to conduct the corresponding *in vivo* experiments. Also, it is interesting to determine the microRNAs targets; it will help to discover their biological functions.

REFERENCES

Bartel D. (2003) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 116–297.

Berezikov E., Guryev V., Belt J., Wienholds E., Plasterk R., Cupen E. (2005) Phylogenetic shadowing and computational identification of human microRNAs genes. *Cell*, **120**, 21–24.

Bonnet E., Wuts J., Rouze P., Van de Peer Y. (2004) Evidence that microRNAs precursors, unlike other non-coding RNAs, have lower folding free energies that random sequences. *Bioinformatics*, **20**, 2911–2917.

Grad Y., Aash Jh., Hayes G.D., Reinhart B.J., Church G.M., Ruvkin G., Kim Jh. (2003) Computational and experimental identification of *C.elegans* microRNAs. *Mol. Cell*, **11**, 1253–1263.

Hofacker I.L., Fekete M., Stadler P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.,* **319**, 1059–1066.

Kong Y., Han J.H. (2005) MicroRNA: biological and computational perspectives. *Geno. Prot. Bioinfo*., **3**, 62–72.

Lim L.P., Lai N.C., Weinstein E.G., Abdelhakim A., Yekta S., Rhoades M.W., Burge C.B., Bartel D.P. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev*., **17**, 991–1008.

Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple alignment aided by quality analysis tools. *Nucl. Acids Res*., **25**, 4876–4882.

TRANSLATION REGULATION IN CHLOROPLASTS

Seliverstov A.V.**, *Lyubetsky V.A.
Institute for Information Transmission Problems, RAS, Moscow, Russia
* Corresponding author: e-mail: slvstv@iitp.ru

Key words: translation, chloroplasts, multiple alignments

SUMMARY

Motivation: Gene expression in chloroplasts of algae and plants is regulated through binding of chloroplast mRNA by nuclear-encoded proteins. It is therefore important to determine such protein binding sites and study them from evolutionary perspective.

Results: An algorithm of finding conservative protein-RNA binding sites is designed, also see details in (Lyubetsky *et al.*, 2004). The algorithm was applied to infer these sites upstream of chloroplast genes. As a result, candidate protein-RNA binding sites were detected upstream of the *atpF*, *petB*, *clpP*, *psaA*, *psbA* and *psbB* genes in many chloroplasts of algae and plants. We suggest that some of these sites are involved in suppressing translation until the completion of splicing.

INTRODUCTION

Gene expression in chloroplasts of algae and plants is regulated by binding of chloroplast mRNA by nuclear-encoded proteins (Nickelsen, 2003). These proteins are involved in editing, translation and maintaining stability of chloroplast mRNA. Detailed analysis of regulatory sites is available from published evidence for alga *Chlamydomonas reinhardtii*, as well as some plants (Hauser *et al*., 1996; Zerges, 2000; Nickelsen, 2003). For example, protein binding to the *psbA* 5′-untranslated region in *C. reinhardtii* results in activation of translation (Hauser *et al*., 1996).

Many chloroplast protein-coding genes contain introns. Thus, their translation should not start immediately after transcription. However, the translation machinery of chloroplasts closely resembles that of bacteria, particularly, in the ribosome being able to immediately follow the RNA-polymerase on mRNA strand. If the ribosome arrives at the end of exon before splicing is completed, the splicing process halts. To avoid this, in some rare cases the AUG start codon is derived from ACG by editing mRNA, which prevents translation from starting immediately (Zerges, 2000). RNA editing is known for chloroplasts of higher plants and is absent, e.g., in the liverwort *Marchantia polymorpha*.

Our algorithm detected candidate protein-RNA binding sites upstream of *atpF*, *petB*, *clpP*, *psaA*, *psbA* and *psbB* genes in many chloroplasts.

We suggest that some of these sites are involved in suppressing translation until splicing is completed. This conjecture is in agreement with observation that multiple alignments of the site-containing regions upstream of these genes are highly conservative, and is also supported by experimental evidence (Hauser *et al*., 1996).

ALGORITHM

Consider a dataset of leader regions upstream of orthologous genes and a corresponding species tree. A set of shallow phylogenetic subtrees (groups of taxa) is selected in the species tree. For each of the groups, the algorithm searches for conserved regions of fixed length n (which can be varied) by finding cliques in a suitable multipartite graph. The basic idea is as follows. The algorithm finds clusters of very similar sites, called signals or motifs, of a fixed length n for each of these phylogenetic groups. From a motif, a weight matrix $4 \times n$ is generated, where the kth column of the matrix, $1 \le k \le n$, contains letter frequencies in the kth site position from the motif. Further, the algorithm generates clusters of weight matrices for different suitable n across all groups. The clusters of matrices are generated accounting for distances in the species tree between the ancestors of the initial phylogenetic groups. The algorithm of clique finding can also be used for constructing these clusters of matrices. In each matrix cluster, the matrices are replaced by the corresponding motif, thus defining sets of motifs. The described procedure can be iterated. The algorithm is described in detail in (Lyubetsky, Seliverstov, 2004).

IMPLEMENTATION AND RESULTS

Chloroplast genomes were obtained from GenBank (NCBI). The initial dataset contained 5'-untranslated intergenic regions from chloroplast genomes of algae and plants.

Occurrence of predicted sites upstream of chloroplast genes is shown in the table.

In many chloroplasts, the algorithm found long conserved binding sites containing conserved helices upstream of the genes *atpF* (ATP-synthetase subunit), *petB* (cytochrome b6), *clpP* (ATP-dependent Clp protease proteolytic subunit), *psaA* (photosystem I P700 apoprotein A1), *psbA* (photosystem II protein D1) and *psbB* (photosystem II P680 chlorophyll A apoprotein).GenBank annotation of the *psbA* gene of *Amborella trichopoda* probably misses a short N-terminal sequence, which might explain why in this case the algorithm failed to find the corresponding motif.

For the genes *atpF*, *clpP* and *petB*, there is a strong correlation between the occurrence of splicing and the existence of the predicted protein-binding sites. On the other hand, with *psaA*, *psbA* and *psbB* no such correlation is found. With *clpP*, *petB*, *psaA*, *psbA*, the sites always contain helices, but for *atpF* and *psbB* they do not.

Table 1. Occurrence of predicted sites and introns upstream of chloroplast genes *atpF*, *petB*, *clpP*, *psaA*, *psbA* and *psbB*. Notation: "+" – candidate protein binding site present; "-" – no candidate binding site; "s" – introns present; "n" – no gene homolog in the species; "&" – helices in the site; "E" – editing of start codon

Species	*atpF*	*clpP*	*petB*	*psaA*	*psbA*	*psbB*
Euglena gracilis	–s	–	-s	–s	-s	-s
Odontella sinensis	–	–	–	+&	+&	–
Guillardia theta	–	–	–	+&	+&	–
Cyanidioschyzon merolae	–	–	–	–	+&	–
Cyanidium caldarium	–	–	–	–	–	–
Porphyra purpurea	–	–	–	+&	+&	+
Gracilaria tenuistipitata	–	–	–	–	+&	–
Chlamydomonas reinhardtii	–	–	–	-s	+&s	–
Nephroselmis olivacea	–	–	–	+&	+&	+
Chaetosphaeridium globosum	–	+&s	-s	+&	+&	+
Mesostigma viride	–	–	–	+&	–	–
Anthoceros formosae	+s	+&s	+&s	+&	+&	+
Marchantia polymorpha	+s	+&s	+&s	+&	+&	+
Huperzia lucidula	+s	+&s	+&s	+&	+&	+

Species	*atpF*	*clpP*	*petB*	*psaA*	*psbA*	*psbB*
Adiantum capillus-veneris	+sE	+&s	-sE	+&	+&	+
Psilotum nudum	+s	+&s	+&s	+&	+&	+
Pinus thunbergii	+s	+&	+&s	+&	+&s	+
Amborella trichopoda	+s	+&s	+&s	+&	–	+
Arabidopsis thaliana	+s	+&s	+&s	+&	+&	+
Atropa belladonna	+s	+&s	+&s	+&	+&	+
Calycanthus floridus	+s	+&s	+&s	+&	+&	+
Cucumis sativus	+s	+&s	+&s	+&	+&	+
Epifagus virginiana	n	+&s	n	n	n	n
Lotus corniculatus	+s	+&s	+&s	+&	+&	+
Nicotiana tabacum	+s	+&s	+&s	+&	+&	+
Nymphaea alba	+s	+&s	+&s	+&	+&	+
Panax ginseng	+s	+&s	+&s	+&	+&	+
Spinacia oleracea	+s	+&s	+&s	+&	+&	+
Oryza nivara, Oryza sativa	+s	+&s	+&s	+&	+&	+
Triticum aestivum	+s	+&s	+&s	+&	+&	+
Zea mays	+s	+&s	+&s	+&	+&	+

DISCUSSION

Conserved motifs detected upstream of the *atpF*, *petB*, *clpP*, *psaA*, *psbA* and *psbB* genes are likely to be involved in translation regulation.

The conserved region upstream of *atpF* contains an AG-rich motif typical for ribosome binding sites, although being considerably longer than typical binding sites. This might be relevant to presence of introns in the gene, which suggests that translation initiates only after completion of splicing.

Upstream region of the *petB* gene does not have a typical ribosome-binding site but instead contains a conserved helix, which might suggest posttranscriptional modification of the 5′-untranslated regions or binding of a translation activator. In all plants, the *petB* gene contains introns.

Translational regulation of the *psbA* gene was experimentally observed in *Chlamydomonas reinhardtii*, where transcription is continuous, but translation is activated at light by a 47 kDa protein that forms a complex with other proteins and mRNA not interacting with mRNA directly (Hauser *et al*., 1996). The complex is inactivated in the dark. The conserved nature of this region in plants and algae might suggest that the translation regulation machinery for gene *psbA* preceded the evolutionary emergence of introns.

Conserved regions in the 5'-untranslated regions of *clpP* and *psbA* genes were observed upstream of almost all their orthologs, even those lacking introns. Notably, conserved RNA motifs in the transcripts of *petB*, *clpP*, *psaA* and *psbA* contain helices with conserved flanks likely interacting with a protein mediator, which is typical for most regulatory systems (Seliverstov *et al*., 2005).

Long conserved motifs were found upstream of the *psaA* and *psbB* genes, which lack introns in all species containing the motifs. On the other hand, in chloroplasts of *Adiantum*, all studied 5′-untranslaled regions are considerably diverged. Hence, the motif was not found upstream of *petB*, while it was in the five other cases. In the latter situation, however, site trees and species trees disagreed considerably at the node containing the name of the corresponding species.

Other intron-containing genes in the studied chloroplast genomes were not found to have conserved 5′-motifs, or their 5′-untranslated regions were too short or absent. Two such examples are discussed. In studied plants, the upstream regions of gene *rbcL* encoding a ribulose 1,5-bisphosphate carboxylase/oxygenase subunit contain only a short conserved motif with the consensus ARGGAGGGACYT, which core constitutes a ribosome-binding site. We have no reason to assign a regulatory role to this motif, as the

rbcL gene in plants lacks introns. On the other hand, *rbcL* contains introns in chloroplasts of both algae *Euglena gracilis* and *Chlamydomonas reinhardtii*, and, in the latter case, it is regulated by mRNA-binding proteins (Hauser *et al*., 1996). This seeming discrepancy is not surprising, since in both algae the structure of 5′-untranslated region is completely different from that in studied plants.

A different situation is with the *ycf3* gene (photosystem I assembly protein Ycf3). It contains introns and a long 5′-untranslated regions not overlapping with other genes in plant chloroplasts, but it was not found to possess conserved motifs.

ACKNOWLEDGEMENTS

The authors are grateful to M.S. Gelfand for valuable discussion and to L.Y. Rusin for discussion and help. This study was partially supported by ISTC 2766.

REFERENCES

Hauser C.R., Gillham N.W., Boynton J.E. (1996) Translation regulation of chloroplast genes. *The J. of Biol. Chemistry*, **271**, 1486–1497.

Lyubetsky V.A., Seliverstov A.V. (2004) Note on cliques and alignments. *Information Processes*, **4**, 241–246.

Nickelsen J. (2003) Chloroplast RNA binding proteins. *Current Genet*, **43**, 392–399.

Seliverstov A.V., Putzer H., Gelfand M.S., Lyubetsky V.A. (2005) Comparative analysis of RNA regulatory elements of amino acid metabolism genes in Actinobacteria. *BMC Microbiology*, **5,** 54.

Zerges W. (2000) Translation in chloroplasts. *Biochimie*, **82**, 583–601.

ALTERNATIVE TRANSCRIPTION WITHIN PROCARYOTIC GENES PREDICTED BY PROMOTER-SEARCH SOFTWARE

Shavkunov K.S.[1], Masulis I.S.[1], Matushkin Yu.G.[2], Ozoline O.N.[*1]
[1] Institute of Cell Biophysics, RAS, Pushchino, Moscow region, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: ozoline@icb.psn.ru

Key words: *Escherichia coli*, alternative transcription, promoters

SUMMARY

Motivation: Mapping of putative promoters within entire genome of *Escherichia coli* by means of pattern-recognition software PlatProm revealed several thousands of sites having high probability to perform promoter function. Along with the expected promoters located upstream from coding sequences PlatProm identified several hundred of very similar signals within coding sequences. Many of them may initiate transcription from the sense strand thus permitting synthesis of shortened mRNA products, not expected *a priory* in bacterial cells.

Results: Here we discuss possible functional significance of intragenic promoters, estimate predictive capacity of our software *in vivo* and *in vitro* and provide experimental evidences that at least one promoter predicted within coding sequences is transcriptionally active.

Availability: Coordinates of predicted transcription start points for alternative transcription are available by request (ozoline@icb.psn.ru).

INTRODUCTION

Genome-wide scanning by PlatProm revealed 709 genes, which have potential internal promoters with a propensity to produce shortened RNA products from the sense strand (Brok-Volchanski *et al*., 2005). At least 46 of them may initiate synthesis of RNAs previously detected in the fraction of small RNAs extracted from bacterial cells (Vogel *et al*., 2003) and considered as products of mRNA degradation. Basically, the presence of intragenic promoters may be required to intensify downstream transcription of neighboring genes (if they have proper orientation) or trap RNA polymerase near real promoters (if they are located at the beginning of the gene (Huerta, Collado-Vides, 2003)). However many predicted promoters lie far from the 5′-end of gene, while the nearest downstream genes have opposite orientation. In these cases internal promoters may be required to express alternative proteins or antisense RNAs to the products of neighboring gene. We, therefore, verified this possibility using available software (ORF Finder and RNA Structure), which allowed identifying open reading frames (ORFs) and characterizing folding propensity of putative RNA product. The scores of the transcription signals found within such genes were compared with known promoters and transcription activity of the promoter, predicted in the middle of the *htgA* gene was verified experimentally.

METHODS AND ALGORITHMS

The search for alternative ORFs was done using ORF Finder (www.ncbi.nlm.nih.gov).

Transcription terminators were found on the basis of the following criteria: 5–10 bp G/C-rich stem, 3–8 bases loop, free energy < -7 kcal/mol, ≥ 4U downstream of the stem (Argaman *et al*., 2001). Folding propensity of potential RNA products was estimated by means of RNA Structure algorithm supplied with thermodynamic scoring system (http://rna.chem.rochester.edu).

Transcription activity of predicted promoters *in vivo* was tested using the total fraction of cellular RNAs isolated from cells taken during exponential and stationary growth phases. cDNA copies of target products were obtained by primer extension using RevertAid M-MuLV reverse transcriptase (Fermentas) and ^{32}P-labeled gene-specific primers. cDNA products were separated from substrates using electrophoresis in 8 % polyacrylamide gel in the presence of 8M urea and visualized by radioautography.

Potassium permanganate footprinting was performed as described (Zaychikov *et al*., 1997). RNA polymerase – promoter complexes were formed at 36 °C in buffer, containing 50 mM Tris-HCl (pH 8.0), 0.1 mM EDTA, 0.1 mM DTT, 10 mM $MgCl_2$, 50 mM NaCl and BSA (5 mg/ml). RNA polymerase was reconstituted from individual subunits, as suggested by Fujita and Ishihama (1996).

RESULTS AND DISCUSSION

PlatProm identified 709 genes containing strong promoter-like signals, from which shortened RNA products potentially may be synthesized. Most of them are expressed as independent transcription units or are the last genes of operons, thus suggesting that some internal promoters may be required to transcribe new genes in intergenic loci or intensify the expression of properly oriented downstream genes. The average size of intergenic regions flanking 3′-ends of such genes is slightly smaller (119 bp) than throughout the whole genome (~150 bp). In general that argues against the first assumption but is in line with the second one. Thus orientations of downstream genes were examined and it was found that two neighboring genes have similar direction in 318 cases.

Transcription signals within remaining 391 genes may be required to synthesize RNAs with a capacity to encode shortened proteins or untranslated RNA products. That is why the sequences located downstream from predicted promoters were further analyzed to reveal alternative ORFs. For this purpose we used nucleotide sequences spanning from predicted promoters to the ends of genes and increased the length of each sequence by 150 bp downstream so as to take into consideration possible terminators located in intergenic regions. Shortened ORFs were found within 305 genes and in 175 cases they are supplied with suitable ribosome binding site (at least 4 matches to AGGAGGT). This set of internal promoters probably has the highest heuristic significance. Fig. 1 represents their scores in comparison with scores of 328 known bacterial promoters, which were absent in compilation used to generate weight matrices of PlatProm. One can see that there are many real promoters with low scores, however in most cases the values of **S** deviate from the background level for more than 3 Std. Surprisingly we found that the distribution of **S** has two well pronounced maxima ($4.5 \le \mathbf{S} \le 5.5$ and $7.5 \le \mathbf{S} \le 9.5$). It could be speculated that corresponding promoters are subjected to different types of regulation. For instance, the set of weaker promoters may require transcription activators for maximal activity, while stronger promoters may be constitutive or their functionality may depend on repressors. In any case, for predictive mapping we used only strong transcription signals (4 Std higher than background level). Fig. 1 demonstrates that distribution of scores for predicted internal promoters also has maximum; positioning of this maximum is the same as in the case of real promoters and there is a fraction of very strong transcription signals (**S** ≥ 13.5, or 6 Std

higher than background). That means that the set of predicted promoters have some features similar with real promoters. All of them can not be ascribed to any known gene.

Fig. 2 exemplifies such situation within gene *htgA*. It encodes positive regulator for promoters recognized by σ^{32} (heat shock regulon) and may be expressed from the σ^{32}-dependent promoter, located 82 bp upstream of the initiating codon of *htgA* (unrecognized by PlatProm) or from the weak σ^{70}-specific promoter, situated 114 bp upstream of ORF. *HtgA* lies between *yaaH* and *yaaI*, transcribed from the opposite strand and fully overlaps with the putative gene *b0011*. At least 3 promoter-like sites are predicted in this region. The strongest one most probably controls transcription of *yaaH*. Others were found within *htgA* and are possibly required to produce both antisense or alternative RNA products in respect to genes *htgA* and *b0011*, as well as an alternative ORF found at the end of *htgA*.

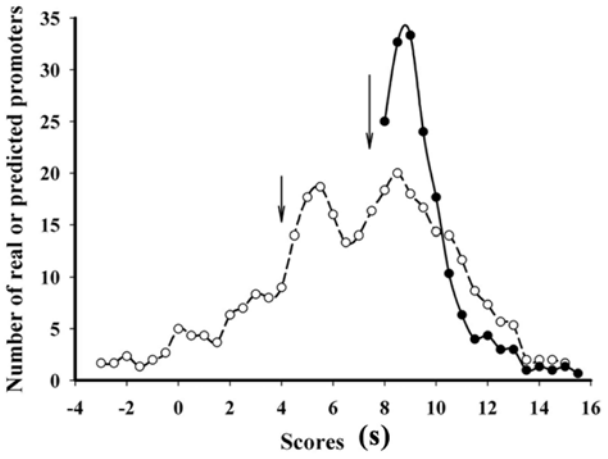


Figure 1. Distribution of scores (S) for 328 bacterial promoters (open circles) and 175 internal promoters (black circles), having a propensity to produce alternative mRNAs. Each point represents the number of nucleotide sequences, having scores within an interval S-0.5 ÷ S. An average S for non-promoter DNAs, estimated by PlatProm was -4.85. Arrows indicate levels of S, which are 3 and 4 standard deviations (Std) larger than this value. Only signals with S ≥ 4 Std were used for predictive mapping. Both curves were smoothened using running window 3.

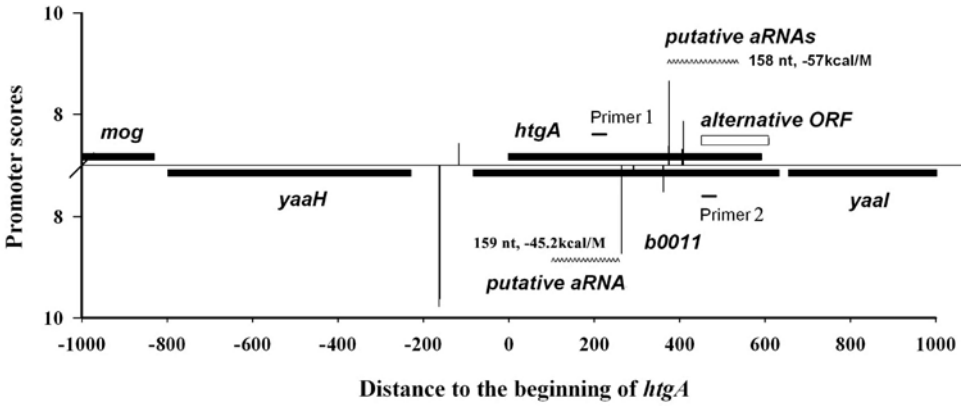


Figure 2. Schematic representation of the chromosome locus, containing gene *htgA*. Solid black lines and arrows drown above or below X axis show positioning of genes in respect to the initiating codon of *htgA* and respective direction of transcription. Bars represent promoters predicted on both strands. Open rectangle indicates location of alternative ORF, while zigzag lines show putative RNA products, which may be synthesized between the predicted promoter and the first r-independent terminator.

Activity of two intrinsic promoters located on the top strand of *htgA* was verified experimentally. They form two clusters and may provide RNA products 92–108 and 62–66 nt long. Primer 2 (Fig. 2) was used for the reaction of reverse transcription to detect the expected RNA products in total fraction of cellular RNAs. They were isolated from cells at exponential and stationary growth phases. At least three short RNAs: 92, 69 and 67 nt long were detected in addition to longer products, originated from upstream promoters (Fig. 3*a*). Their abundance does not depend on the growth phase. Although sizes of these products are very close to the expected ones, some of them may be products of mRNA degradation. That is why we used the potassium permanganate footprinting technique to answer the question whether RNA polymerase forms open promoter complexes in these regions (Fig. 3*b*).

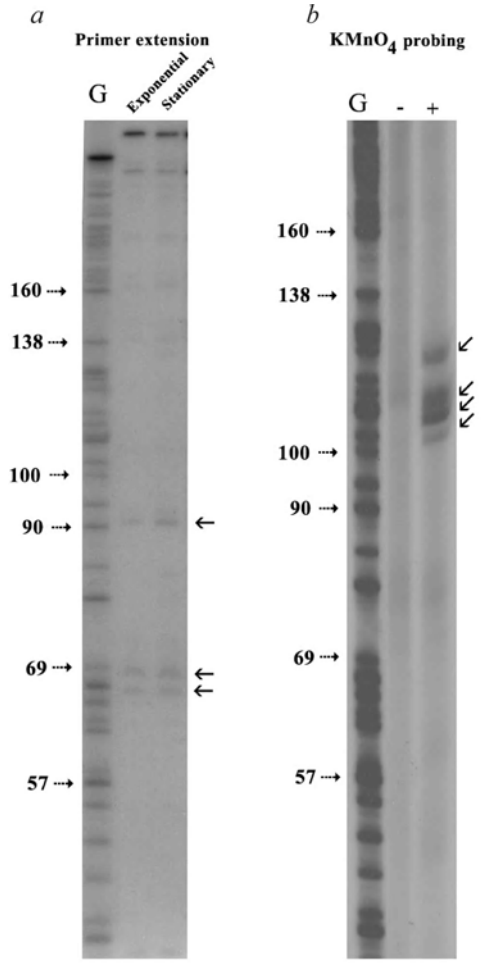


Figure 3. Experimental verification of promoter activity by means of primer extension (*a*) and potassium permanganate footprinting (*b*). (*a*): Primer 2 and 1 ng of cellular RNA were used to obtain cDNA copies. cDNA products were separated on polyacrylamide gel (8 %) and visualized by radioautography. Arrows on the right indicate observed shortened RNAs. (*b*): PCR amplified DNA fragment (Primers 1 and 2 shown in Fig. 1.) was used to study an RNA polymerase binding capacity to predicted promoters. Complexes were formed as described in Methods and Algorithms. RNA polymerase – promoter ratio was 1:4(M:M). Marks "-" and "+" denote samples, containing free DNA fragment and DNA-protein complexes, respectively. Arrows on the right indicate bands representing the specific modification of unpaired thymines. Both gels were calibrated by standard G-specific ladder of another DNA fragment. Ciphers on the left reflect the sizes of indicated fragments.

Modifying only unpaired thymines, the potassium permanganate provides an excellent opportunity to reveal transcriptional bubble and, therefore, to localize specific RNA-polymerase binding site(s) on DNA. The data obtained clearly indicate that *in vitro* open complexes are really formed and the binding site is located near the cluster of predicted transcription start points with high scores (92–108 bp upstream of primer 2). There are no any reactive thymines near weaker transcription signals (62–66 bp far from the primer) thus indicating that RNA polymerase selects more strong promoter site, while two products detected in the reaction of primer extension may result from RNA decay. Fig. 4 shows the nucleotide sequence of the region containing active promoter.

The observed pattern of reactivity against potassium permanganate allows a possibility that RNA polymerase can initiate transcription from all three predicted start points in the cluster (genomic coordinates are: 11090, 11099 and 11102 on the + strand). The major product observed *in vivo* (Fig. 3*a*) is, however 6–3 nt shorter than expected in the case if RNAs are initiated from promoter having almost perfect -35 and -10 elements (98 and 95 bp from primer), and more pronounced transcription bubble. Weaker upstream promoter with a capacity to give 107 nt RNA (initiated from position 11090) also binds RNA polymerase; although *in vivo* the product of exactly this length was observed only upon longer exposition. In any case, the internal promoter predicted within gene *htgA* is active. RNA transcribed from this promoter may encode a 31 amino acids long polypeptide, with ORF shifted on 1 position in respect to mRNA of *htgA*. This product has no sequence homology with any other known protein. Alternatively 158 nt long RNA transcribed between the verified promoter and the first ρ-independent terminator may function as antisense RNA to mRNA of hypothetical protein *b0011*. Free energy of folding for this transcript (-57 kcal/M) is typical for small regulatory RNAs of this length.

Figure 4. Nucleotide sequence of the predicted internal promoter. Lower case letters indicate predicted start points of transcription. Suitable -35 regions are underlined; -10 regions are shown by larger font letters. Adenines complementary to thymines, modified by potassium permanganate are double underlined. Ciphers above the sequence indicate the expected length of the product originated from marked position.

Taken together we conclude that PlatProm may be used as a tool predicting novel transcripts in the genome of *E. coli*.

ACKNOWLEDGEMENTS

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

REFERENCES

Argaman L., Hershberg R., Vogel J., Bejerano G., Wagner E.G.H., Margalit H., Altuvia S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol*., **11**, 941–950.

Brok-Volchanski A.S., Masulis I.S., Shavkunov K.S., Lukyanov V.I., Purtov Yu.A., Kostyanicina E.G., Deev A.A., Ozoline O.N. (2005) Predicting sRNA genes in the genome of *E. coli* by the promoter-search algorithm PlatProm. In Kolchanov N., Hofestaedt R., Milanesi L., (eds), *Bioinformatics of Genome Regulation and Structure II*. Springer, pp. 11–20.

Fujita N., Ishihama A. (1996) Reconstitution of RNA polymerase. *Meth. Enzymol*., **273**, 121–130.

Huerta A.M., Collado-Vides J. (2003) Sigma-70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol*., **333**, 261–278.

Vogel J., Bartels V., Tang T.H., Churakov G., Slagter-Jager J.G., Huttenhofer A., Wagner E.G.H. (2003) Rnomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucl. Acids Res*., **31**, 6435–6443.

Zaychikov E.**,** Denissova L.**,** Meier T.**,** Götte M., Heumann H. (1997) Influence of Mg^{2+} and temperature on formation of the transcription bubble. *J. Biol. Chem*., **272**(4), 2259–2267.

ANALYSIS OF REGULATORY REGION OF *BACILLUS INTERMEDIUS* GLUTAMYL ENDOPEPTIDASE GENE

Shagimardanova E.I.**, Shamsutdinov T.R., Chastuchina I.B., Sharipova M.R.
Kazan State University, Kazan, Russia
* Corresponding author: e-mail: rjuka@mail.ru

Key words: glutamyl endopeptidase, gene expression, consensus sequence, catabolite repression, phosphorelay

SUMMARY

Motivation: Representatives of the *Bacillus* genera are nonpathogenic and are suitable for producing various proteases. The cloning and sequencing of genes give the opportunity to predict the role of the proteins in the cells, its cooperation with other molecules and ways of their own regulation.

Results: The analysis of glutamyl endopeptidase promoter region was initially developed. The presence of consensus sequences for binding with CcpA, AbrB, Spo0A regulatory proteins were shown.

INTRODUCTION

Many species of the genus *Bacillus* produce a variety of extracellular and intracellular proteases. Extracellular enzymes were extensively studied due to its commercial importance in the fields of medicine and household chemical goods. The investigation of new group of proteolytic enzymes designated as glutamyl endopeptidases begins after discovery in 1972 serine protease from strain V8 *Staphylococcus aureus* (Drapeau *et al*., 1972). This enzymes posses narrow substrate specificity and split only the peptide bonds formed by α-carboxyl groups of glutamic and aspartic acids (Rudenskaya, 1998). At present more that 100 glutamyl endopeptidases have been assigned to the subfamily within the chimotrypsin family of serine proteases. All the enzymes belonging to this subfamily have in common the catalytic domain, characterized by the presence of "structurally conserved regions". These are secreted proteins of 18–29 kDa, their pI varying in a wide range of pH values. Interesting particularity of these enzymes is the presence of one optimal pH value while hydrolyzing peptide substrates and two pH-optimums while hydrolyzing protein substrates (Rudenskaya, 1998). Glutamyl endopeptidases demonstrate vary high resistance to the different inhibitors. Thus, enzymatic properties and structure of bacterial glutamyl endopeptidases are well enough studied, whereas their biological role is still unclear and too little is currently known about the mechanisms of biogenesis of these enzymes. Therefore, further research of the biosynthesis of bacterial glutamyl endopeptidases would be desirable.

Glutamyl endopeptidase from streptomycin-resistant strain *B. intermedius* 3–19 (BIEP) was isolated and characterized (Leshchinskaya *et al*., 1997). The gene encoding for *B. intermedius* glutamyl endopeptidase was cloned and sequenced (EMBL accession number Y15136) (Rebrikov *et al*., 1999). However, the mechanisms involved in the regulation of glutamyl endopeptidase gene expression remains unclear. Here we report the potential mechanisms controlled the expression of the *B. intermedius* glutamyl endopeptidase gene.

METHODS AND ALGORITHMS

Nucleotide sequence of the gene for *B. intermedius* glutamyl endopeptidase analyzed in this study is available in the EMBL database (EMBL accession number Y15136). The DNA sequence preceding the gene for glutamyl endopeptidase was inspected for the occurrence of the characteristic –35 and –10 boxes of SigA-type promoters (Helmann, 1995) by using the Softberry PROM (Prediction of Bacterial Promoters) network server (http://www.softberry.com). The consensus nucleotide sequences were detected using Vector NTI version 8 software.

IMPLEMENTATIONS AND RESULTS

To elucidate the mechanisms of *gseBi* expression the analysis of its promoter region was initially developed. The promoter region of *B. intermedius* glutamyl endopeptidase gene is shown in Fig. 1. Potential –10 and –35 regions for recognition by sigma A (75 % of homology) identified by using Softberry BPROM network server are underlined.

```
  1 cagctgccaccttcatctgaagtaccgatgatcatgatcggaccgggtacaggaatcgct
 61 ccatttagagggtttgttcaggcaagagaagtgtggcagaaggaaggaaanccactaggc
121 gaagctcacctgtattttggctgtcgtcaccctctTGAAGATgatctgtatttgagGAA
181 atgcagcttgcagcgcaaaaaggagttgttcacatccaccgggcttattctcgtcacaaa
241 gagcaaaaagtatatgtccagcattgttgaaagaagacggcggcatgctgatcaagtta
301 cttgaccaaggtgggtatctttacgtgtgcggggacggaaaagttatggcaccagacgta
361 gaggctacgcttatcgacctctatcaaaacgagaaacattgctcgaaggaaacagctgaa
421 aattggctgacaacccttgcaaatgacaatagatatgtaaaagatgtgtggagctgaaaa
481 attaggaagaccgccaattaggcggtttttctatttcatagagagagatcaatagaatg
541 aaggttggaaagatacaaaacacctaatttaaaaatgaaatattttgtaaaaaataagaa
              -35                    -10                  SD
601 tattctctcatttactccaatatgaaacaatcgtatgatttttgatataggacataaagg
661 aggaatatg
```

Figure 1. The nucleotide sequence of gseBi gene. The position of the Shine-Dalgarno consensus sequence and the -35 and -10 regions are underlined. Potential Spo0A binding sites are shown in bold and underlined. A regions showing homology to the consensus sequences for site binding the catabolite repressor, TGWAARCGYTWNCW are boxed and to the AbrB regulatory protein, WAWWTTTWCAAAAAAW is shaded.

The expression of genes, participating in sporulation, is controlled by two-component signal transduction system KinA/Spo0F/Spo0A. For such genes there is a consensus nucleotide sequence (TGNCGAA) for binding with the transcription factor Spo0A. Spo0A can serve either as a repressor or an activator of transcription, depending on the target gene (Strauch *et al*., 1990; Burbulys *et al*., 1991). In the regulatory region of the gene for glutamyl endopeptidase the nucleotide sequence sharing 86 % identity with the proposed consensus sequence for binding with Spo0A regulatory protein was identified. These sequences appeared to be organized as direct tandem repeats.

The regulatory region of *gseBi* gene also contains four sequences with structural homology 86 % to specific target site for (WAWWTTTWCAAAAAAW) (Strauch, 1995) for binding with pleiotropic repressor of the early sporulation genes, AbrB (-200-215). The promoter region of *gseBi* was screened for homology with the sequence, TGWNANCGNTNWCA, the consensus operator sequence for binding of the catabolite repressor CcpA (LeDeaux *et al*., 1997). Regions with 71–78 % homology on nucleotide position -88-101, -159-172, -206-219 downstream of the transcriptional start point were found.

DISCUSSION

In natural environment microorganisms are always subjected to a variety of stresses and nutrient deprivation. In this period cells induce the production of biological-active molecules, including different enzymes such as proteases.

Extracellular Glu-endopeptidase from *B. intermedius* is exerted at the late phase of bacterial growth. The biosynthesis of BIGEP as well as other *Bacillus* glutamyl endopeptidases is enhanced before sporulation (Gabdrakhmanova *et al*., 1999). This observation allows to suggest the possibility of participation of Spo0A-phosphorelay in control of glu-endopeptidase synthesis. The promoter region of this enzyme was shown to have the potential sites to binding with Spo0A-protein (Fig. 1). The organization this sequences as direct tandem repeats may enhance the frequency of binding with regulatory protein. This suggestion confirms with experimental data: in recombinant *B. subtilis* strain with inactivated Spo0A production of the glutamyl endopeptidase decreased 1.5-fold. Thus, found sites appeared to be usable, but this sequences not the only way to regulate the glutamyl endopeptidase gene expression.

There is also the nucleotide sequence for interaction with AbrB protein. AbrB is the pleiotropic repressor of early sporulation genes. Results obtained by practical consideration demonstrated that the level of glutamyl endopeptidase gene expression in the strain carrying *abrB* mutation was higher than that in control suggesting possible binding of AbrB with promoter region of the *gse Bi* and the following negative control.

As is generally known, many degradation enzymes are under control of catabolite repression mechanism. CcpA, a member of the family of transcriptional regulators, is believed to be central to catabolite repression of many catabolic operons in gram-positive bacteria (Henkin, 1996). CcpA protein binds specifically to the *cre* (catabolite-responsive element) sequence in the target genes, preventing transcription initiation in the presence of glucose. In promoter region of *gse Bi* nucleotide sequence sharing 78 % identity with *cre* sequence was found, and, besides, several more sequences with 71 % homology with this regulatory element were identified (Fig. 1). The identification of *cre*-consensus sequence in *gse*Bi gene promoter in conjunction with the experimental researches suggest the involvement of carbon catabolite repression in the expression of *gseBi.* But, in fact, the decrease of enzymatic activity in the presence of preferred carbon source occurs only at the early stationary phase. In our experimental study, we recorded no effect of glucose on glutamyl endopeptidase production during the late stationary phase. That is in accordance with literature data: transcription of the *ccpA* gene is mediated by the vegetative σ^A –factor, and during the late stationary phase, when sigma factors involved in sporulation are activated, *ccpA* expression decreases substantially. Thus, this data indicate the involvement of potential CcpA consensus sequence in the expression of *gse Bi* during the vegetative growth and its silence at the late stationary phase.

The monitoring sites and experimental data suggest that two-component regulatory system Spo0F/Spo0A and global regulatory network including catabolite repression (CcpA) and AbrB-regulation would be involved in the control of *gse Bi* expression.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation of Basic Research (grant 05-04-48182-a).

REFERENCES

Burbulys D. *et al.* (1991) Initiation of sporulation in B. subtilis is controlled by a multicomponent phosphorelay. *Cell*, **64**, 542–552.

Drapeau G.R. *et al*. (1972) Purification and properties of an extracellular protease of *Staphylococcus aureus. J. Biol. Chem*., **247**, 6720–6726.

Gabdrakhmanova L.A. *et al.* (1999) Biosynthesis and localization of glutamyl endopeptidase from *Bacillus intermedius* 3–19. *Microbios*., **100**, 97–108.

Helmann J.D. (1995) Compilation and analysis of *Bacillus subtilis* σA-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA. *Nucl. Acids Res*., **23**, 2351–2360.

Henkin T.M. (1996) The role of CcpA transcriptional regulator in carbon metabolism in *Bacillus subtilis*. *FEMS Microbiol. Lett*., **135**, 9–15.

LeDeaux *et al.* (1997) Analysis of non-polar deletion mutations in the genes of the spo0K (opp) operon of *Bacillus subtilis*. *FEMS Microbiol. Lett*., **153**, 63–69.

Leshchinskaya I.B. *et al.* (1997) Glutamyl endopeptidase of *Bacillus intermedius*, strain 3–19. *FEBS Lett*., **404**, 241–244.

Rebrikov D.V. *et al*. (1999) Molecular cloning and nucleotide sequence of *Bacillus intermedius* glutamyl endopeptidase gene. *J. Prot. Chem*., **18**, 21–26.

Rudenskaya G.N. (1998) Glutamyl endopeotidases from microorganisms – a new subfamily of chymotrypsin proteinases. *Bioorg. Chem*., **24**, 256–261.

Strauch M.A. *et al*. (1990) The SpoOA protein of *Bacillus subtilis* is a repressor of the *abrB* gene. *Proc. Natl. Acad. Sci. USA*, **85**, 1801–1805.

Strauch M.A. (1995) Delineation of AbrB-binding sites on the *Bacillus subtilis spo0H, ftsAZ* and *pbpE* promoters and use of a derived homology to identify a previously unsuspected binding site in the *bsuBl* methylase promoter. *J. Bacteriol*., **177**, 6999–7002.

PROMOTER MODELING APPROACHES APPLIED TO THE INVESTIGATION OF p63 UP- AND DOWNSTREAM PROMOTERS

***Shelest E.S.*[*1], *Wingender E.*[1,2]**
[1] Dept. of Bioinformatics, UKG, University of Göttingen, Goldschmidtstr. 1, D-37077 Göttingen;
[2] BIOBASE GmbH, Halchtersche Str. 33, D-38304 Wolfenbüttel, Germany
[*] Corresponding author: e-mail: ksh@bioinf.med.uni-goettingen.de

Key words: p63, promoter modeling, transcription factor binding sites, keratinocytes

SUMMARY

Motivation: The transcription factor p63 is a homolog of p53, the tumor suppressor in higher mammals. p63, but not p53, can be expressed from at least two alternative transcription start sites (TSS), yielding a full-length form from the upstream and truncated form (ΔN) from the downstream TSS. The ΔN form acts as p53 antagonist, which makes the understanding of its regulation an important task. To date, the regulators of the p63 promoters are still to be identified.

Results: Comparative analysis of the p63 promoter regions of several species revealed highly conserved combinations of transcription factor binding sites (TFBS), which are suggested as the models of the regulatory patterns for the up- and downstream promoters. The predicted involvement of RXR in the regulation of p63 is in agreement with the experimental data. The other predictions are presently under the experimental evaluation.

With this work, we demonstrate the applicability of the methods of promoter modeling previously developed in our group to a new kind of task: investigation of unknown regulatory patterns in a single promoter based on phylogenetic comparisons.

INTRODUCTION

p53, p63, and p73 constitute a family of DNA-binding proteins that share significant sequence homology. Both p63 and p73 can be expressed from at least two alternative transcription start sites (TSS), yielding full-length forms (transactivating, TA) from the upstream and truncated forms (ΔN) from the downstream TSS. The up- and downstream TSS are under control of two distinct promoters. In spite of the structural similarity, p63 and p73 demonstrate functional differences from p53 and between each other (Waltermann *et al.*, 2003). Compared to p53, the roles of its homologs are more diverse. Although the TA forms of the factors can induce p53-responsive genes, the factors are not specifically assigned to tumor cells and cannot be defined as tumor-suppressors. p73 is frequently overexpressed in various malignancies, but also plays role in normal development of nervous and immune systems; p63 is known to be important for skin development, being specifically expressed in keratinocytes. Thus, the diversity of the functions and complexity of the transcription model makes the understanding of the transcription regulation of p63 and p73 a challenging task.

Little is known about the regulation of transcription of the two homologs of p53. It has been shown that the p73 promoters are regulated by E2F, p53 and (indirectly) by TGFβ (transforming growth factor β), but these factors have either no or negative effect on the

p63 promoter. To date, there is no positive information about the p63 regulators. Thus, we decided to undertake a computational analysis of the promoters in order to supply the experimentalists with suggestions, which they could later confirm with their methods.

METHODS AND ALGORITHMS

Promoter sequences were extracted from ENSEMBL based on sequence homology to the corresponding regions in the human genome (AB055067 for deltaNp73 and AF124530 for dNp63). Homologous sequences were identified with NCBI Blast (http://www.ncbi.nlm.nih.gov). The set of the p63 upstream promoters contained 5 sequences (human, mouse, rat, cow, and dog). The downstream promoters were represented by 6 species for the (human, mouse, rat, cow, dog, and chicken). The length of the sequences in both sets was 1500bp (-1399/+100).

Negative training set consisted of all human promoter sequences from EPD database (1871 seq.). The set was checked for the absence of p63 promoters. The length of the sequences was 1500bp (-1399/+100).

Multiple alignments of orthologous promoter sequences were performed with the Multi-LAGAN tool (http://lagan.stanford.edu/lagan_web/index.shtml)

Search for potential binding sites was undertaken with the help of the Match^{TM} tool (Kel *et al*., 2003) (http://www.biobase.de/cgi-bin/biobase/transfac/start.cgi). The thresholds for the matrix search were adjusted in such a way that the matrix (or set of matrices) for each factor could re-identify 80% of the true positive set (i.e., the set of genuine binding sites). The binding sites for p63 (p53) were predicted with a tool P53MH (Hoh *et al*., 2002), which searches for two p53 binding sites separated by a gap up to 13 bp.

The prediction of potentially functional TFBS pairs was performed by two independent methods: (i) set of approaches to promoter model construction as described in (Shelest, Wingender, 2005); the approach of distance distributions described in (Shelest, 2006).

Databases

Eukaryotic Promoter Database (http://www.epd.isb-sib.ch), release 77-1

Ensembl Genome Browser (http://www.ensembl.org/index.html)

TRANSFAC® Professional release 9.4 (http://www.biobase.de)

RESULTS AND DISCUSSION

The starting point of the analysis was the fact that the transcription of p63 is keratinocyte-specific; thus, it was reasonable to check in the first place the transcription factors active in these cells. The search for the “keratinocytes” in the field "cell specificity" in TRANSFAC database revealed 9 transcription factors (AP-2, Sp1, KRF-1, RXR-α, ESE-2, ESE-2b, POU2F3, ΔNp63, p63), from which only 6 possessed PWMs from the TRANSFAC matrix library. These 6 factors (AP-2, Sp1, RXR-α, POU2F3 (Oct-2), ΔNp63, p63) were taken for the analysis (ΔNp63 and p63 have the binding sites identical to p53, hence the same matrix).

After the identification of single potential binding sites with the help of the Match^{TM} tool, we analyzed the occurrences of combinations of these TFBS. Up- and downstream promoters were considered separately.

The predictions for the TF binding site (TFBS) combinations were made by two independent approaches. The first (Shelest, Wingender, 2005) considers overrepresentation of sequences containing certain TFBS pairs in the investigated set compared with a negative set. We adjusted the parameters in such a way that the pairs were present in 100 % of the investigated sets, and less than in 10 % of the negative training set. The results are shown in Table 1, right column. The second approach, called

"distance distributions approach" (Shelest, 2006), considers the pairs which occur on "overrepresented" distances in comparison with analytically calculated profile of distance distribution in the random case (i.e., when the binding sites are distributed randomly). The results of the application of this method are shown in the left column of the Table 1. Note that both approaches identified practically the same pairs (Table 1).

On the next step of the analysis we looked whether the found combinations of TFBS were evolutionary conserved. For that we mapped the TFBS on the plots representing the conserved regions of the promoters (Fig. 1 and 2).

Table 1. Comparison of the predictions of TFBS pairs made by the two approaches

Distance distributions approach	TFBS pairs approach
A. DOWNSTREAM PROMOTERS	
AP-2 – p53 (97-101)	
Oct-2-Oct-2-(34-41)	**Oct-2-Oct-2 (39-69)**
Oct-2 - Sp1 (84)	**Oct-2-Sp1 (84)**
Oct-2-RXR (12-15)	**Oct-2-RXR (15)**
RXR-RXR (30)	**RXR-RXR (30)**
Sp1-p53 (23-34)	Sp1-RXR (39), (45) and (69)
Sp1-Sp1 (38-52) and **(126-135)**	**Sp1-Sp1 (126-130)**
B. UPSTREAM PROMOTERS	
AP-2 – Sp1 (26), (32), **(81-82)**	**AP-2 – Sp1 (80-82)**
AP-2-Oct-2 (52), (121-125)	**AP-2-Oct-2 (52)**
AP-2-RXR (78-**81**)	**AP-2-RXR (81)**
Oct-Oct (30-31)	
Oct-2 - Sp1 (69-102)	**Oct-2 - Sp1 (78-84)**
Oct-2-RXR (7-10)	**Oct-2-RXR (7-8)**
Sp1-RXR (12), **(72-82)**	**Sp1-RXR (81)**
Sp1-Sp1 (5)	

Notes. Marked with bold font are the coinciding pairs. In parentheses is shown the distance range (i.e., not less than the first number, not more than the last).

As it can be seen on the Fig. 1, the highly conserved regulatory module is constituted from TFBS for 3 factors: Oct-2 (POU2F3), RXR and Sp1. RXR sites can be used by retinoic acid, the involvement of which is in agreement with previously reported experimental data (Bamberger *et al*., 2002).

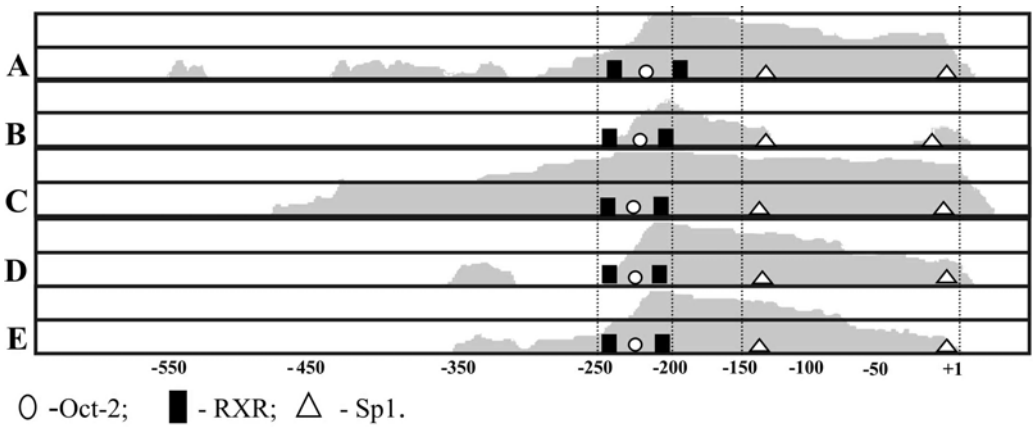


Figure 1. Distribution of the TF binding sites in the regions conserved between the human p63 downstream promoter sequence and five orthologs: cow (A), chicken (B), dog (C), mouse(D) and rat (E).

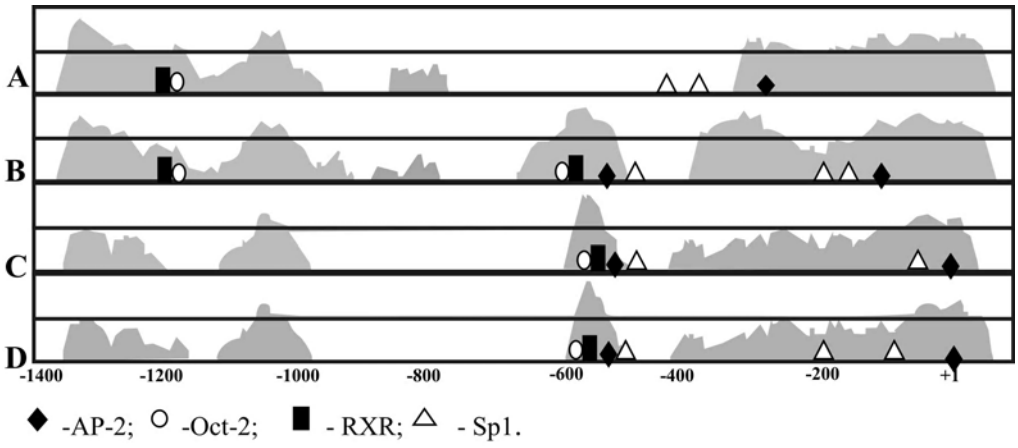


Figure 2. Distribution of the TF binding sites in the regions conserved between the human p63 upstream promoter sequence and four orthologs: cow (A), dog (B), mouse (C) and rat (D).

The distribution of the sites in the upstream promoters (Fig. 2) deserves more discussion. One can notice that there are three "islands" of high conservation: –1380 - –1000 (worse conserved between rodents and human), -600 - -500 (not present in cow) and –400 - +80. An interesting behavior demonstrates the combination of RXR-Oct-2 TFBS. It is conserved in the region –1380 - –1000 in cow and dog, but is not present in mouse and rat where this region is also not conserved. However, it appears now in the region –600- -400, in which it is not found in cow, but is detected in other species, appearing twice in dog. We can speculate that this combination was present in two copies in the common ancestors of these species and retained as such in the dog and human, whereas cow has lost one and rodents the other copy.

The predicted combinations are presently under experimental verification in the laboratory of Molecular Oncology headed by Prof. M. Dobbelstein (Göttingen).

ACKNOWLEDGEMENTS

The work was supported by grant from the German Federal Ministry of Education and Research to the Intergenomics Bioinformatics Competence Center (grant No. 031U110A).

REFERENCES

Bamberger C., Pollet D., Schmale H. (2002) Retinoic acid inhibits downregulation of DeltaNp63alpha expression during terminal differentiation of human primary keratinocytes. *J. Invest. Dermatol.*, **1**, 133–138.

Hoh J. *et al.* (2002) The p53MH algorithm and its application in detecting p53-responsive elements. *Proc. Natl. Acad. Sci. USA*, **99**(13), 8467–8472.

Kel A. *et al.* (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res.*, **31**, 3576–3579.

Shelest E. (2006) Genetic network of antibacterial responses of eukaryotic cells. Bioinformatics analysis and modeling. *Dissertation*, Technical University, Braunschweig, 2006 (http://opus.tu-bs.de/opus/volltexte/2006/845/pdf/Thesis.pdf).

Shelest E., Wingender E. (2005) Construction of predictive promoter models on the example of antibacterial response of human epithelial cells. *Theor. Biol. Med. Model.*, **2**, 2.

Waltermann A., Kartasheva N.N., Dobbelstein M. (2003) Differential regulation of p63 and p73 expression. *Oncogene*, **22**, 5686–5693.

A STEP BEYOND PLANT TRANSCRIPT'S POLYADENYLATION SITE

Smetanin D.V.**, *Chumak N.M.
Krasnodar Research Institute of Agriculture, Krasnodar, Russia
* Corresponding author: e-mail: molbio@mail.kubtelecom.ru

Key words: 3′UTR, polyadenylation site, mRNA, EST, computer analysis

SUMMARY

Motivation: Variation in polyadenylation rates affect mRNA stability, translation and transport and could be mediated through variation of control sequences. Most plant transcripts possess multiple polyadenylation sites, and therefore it is possible to search for downstream elements of poly(A) signal (i.e. part of signal beyond polyadenylation/cleavage site) without usage genomic template, that extremely extend a range of organisms which can be analyzed.

Results: We establish a special approach to search poly(A) signals around polyadenylation/cleavage site of transcript that based on EST analysis only. We applied this method to demonstration the structure of poly(A) signal in *Oryza sativa* and *Triticum aestivum*.

Availability: Analysis pipeline, implemented as a set of Perl scripts and processed datasets available from authors by request.

INTRODUCTION

The 3` ends of most processed eukaryotic mRNAs have a poly(A) tail. Variation in polyadenylation rates affect mRNA stability, translation and transport and could be mediated through variation of control sequences.

The polyadenylation process requires two major components: the cis-elements or poly(A) signals of the pre-mRNA, and the trans-acting factors that carry out the cleavage and addition of the poly(A) tail at the 3′-end.

Primary information about the poly(A) signal elements was derived mostly through conventional genetic and some biochemical analysis. The availability of full sequenced genomes and expressed sequence tags (EST) provides an abundant resource for analysis of transcripts. Moreover, since most ESTs are primed from 3′ termini of mRNA, the EST resource is particularly enriched in final 3′UTR sequences and it is possible to search for poly(A) signals using bioinformatics tools. The efficiency of this approach has been proved by many publications revisiting the poly(A) cis-elements in different organisms (Brockman *et al*., 2005).

But all this studies were conducted on organisms with full sequenced genomes. If it was not a case, studies were limited by sequence analysis until polyadenylation site only.

We used a special approach to search for downstream poly(A) signals i. e. signals beyond polyadenylation/cleavage site of transcript. This approach gets possibility to use ESTs only and does not require genomic sequences.

METHODS AND ALGORITHMS

We used EST datasets of *Oryza sativa* and *Triticum aestivum,* from dbEST (release 10.20.05). The input EST datasets were cleaned up to remove contaminating sequences (vectors, adapter *et al*.). The ESTs with a poly(A) or poly(T) extremity of length 10 or more (exact Perl rules: /A{5,}.{0,1}N*A{5, }.{0,20}$/ or /^.{0,20}T{5, }.{0,1}N*T{5, }/) were retained, poly(A) and poly(T) stretches were removed and sequences were oriented in one direction. Known repeats were masked using Repeat Masker. Sequences were truncated till 500 nt parts, adjoining detected terminal poly(A) stretch and after that TGICL (Pertea *et al*., 2003) was run independently for set of sequences of each species. ESTs from every TGICL cluster (actually CAP3 contig) were aligned based on CAP3 data and poly(A) sites distribution for cluster (i. e. hypothetical mRNA) were defined.

Our approach for search cis-elements of poly(A) signals is based on the analysis of alignments of ESTs aligned with regard to the position of cleavage site. First, ESTs inside each cluster were aligned base on common sequence, and then sequence from the current CS till the end of the longest EST from the cluster was regarded as region beyond CS for current site. The same procedure was applied for each next site, except the most distal one, for which information about region beyond CS was not possible to get (Fig. 1).

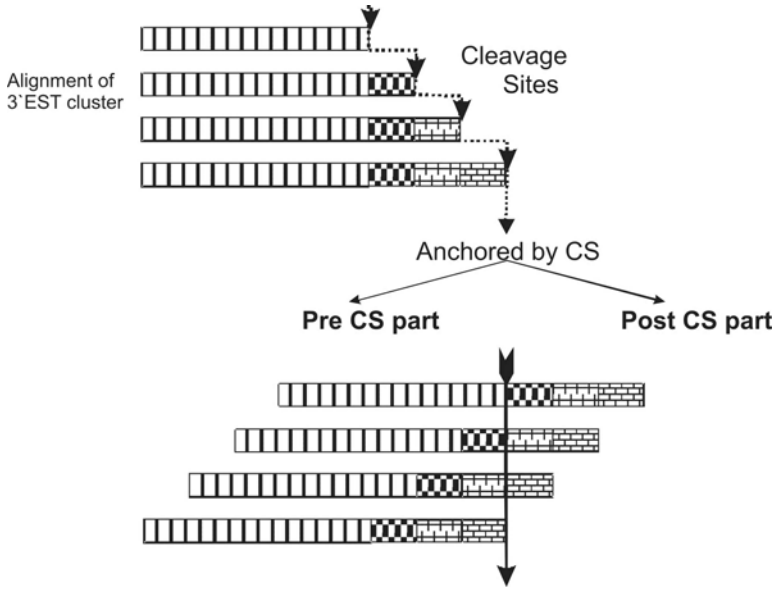


Figure 1. Schema of extended alignment construction for cluster of EST.
CS – cleavage/polyadenylation site.

Second, all extended by such manner ESTs were aligned altogether with regard to the position of cleavage site and the position dependent occurrence frequencies of the hexanucleotides (words) were determined. For single nucleotide composition analysis we used two different alignments one for pre- and another for post-cleavage site sequences.

In order to retrieve genomic regions around each polyadenylation site (in this case we considered sites when at least two ESTs finished at the same position). ESTs formed valid polyadenylation sites where aligned on the rice genome (IRGSP Release Build 4.0) with the BLAT program (Kent, 2002). We retained alignments meeting the following criteria: contain final 3′ part of EST; length > 60 nt; E-value < 0.001; identity > 98 %; no dangling end in 3′ EST direction. For each polyadenylation site the +/- 300 nt region was extracted. Redundant sequences were eliminated.

For word count we used overlapping windows and counted all possible six-letter words in each nucleotide position of -35/+15 around each polyadenylation site. The Z-score was computed based on the first-order Markov chain model of the same region, via the rmes.gaussien application (Schbath, 1997). Hexamers with Z-score above 3 were selected for further analysis.

Analysis pipeline implemented as a set of Perl scripts and use Bioperl and PostgreSQL. A statistical analysis was done in R.

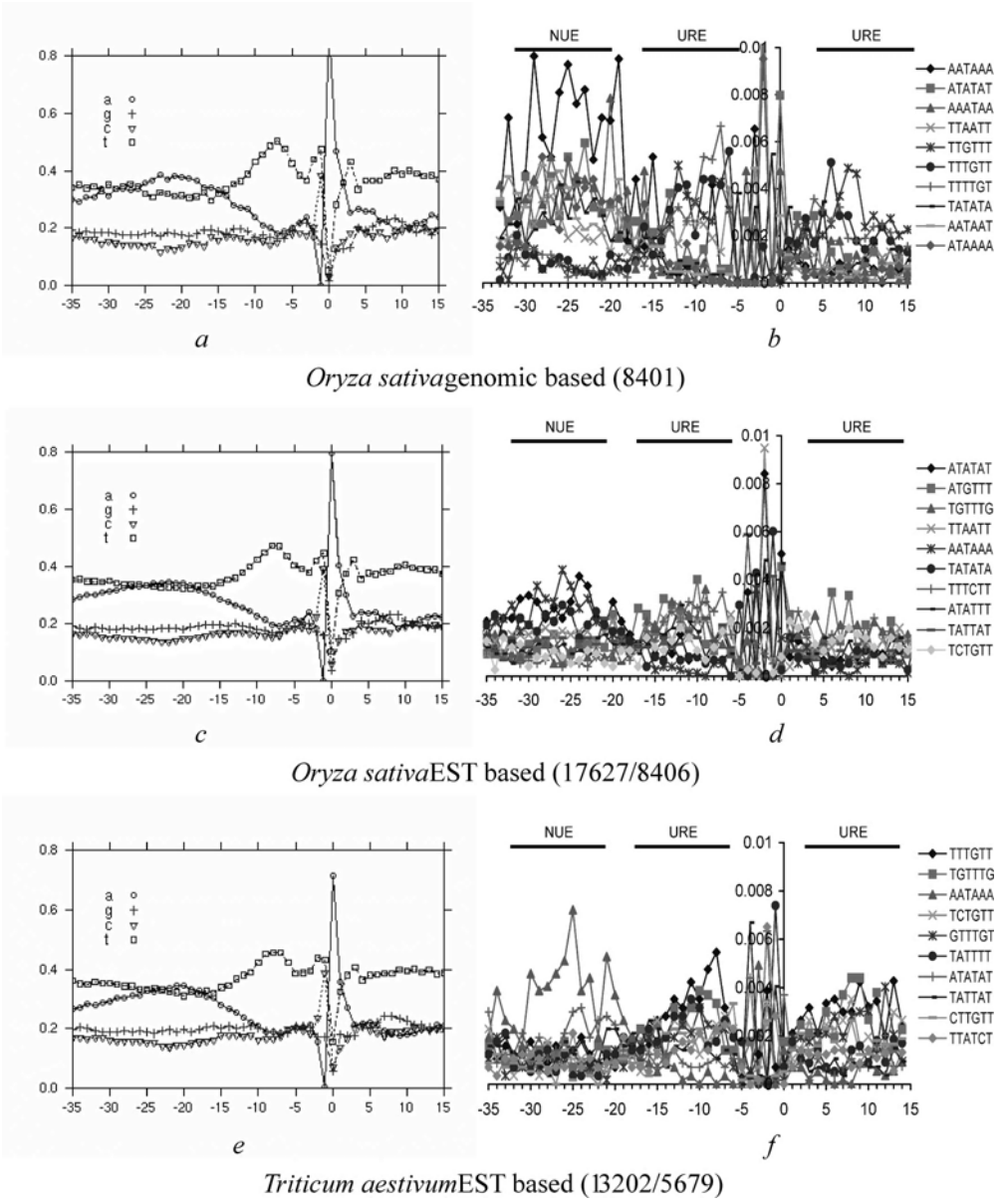


Figure 2. Single-nucleotide frequencies of sequences (*a*, *c*, *e*) and positional distribution of top 10 most frequent 6-nt words (*b*, *d*, *f*) around polyadenylation/cleavage site of plant pre-mRNAs. The species names are marked under each graph, the number of sequences used for each graph is shown in parentheses. The y-axis for *a*, *c*, *e* is fractional abundance of bases and for *b*, *d*, *f* is fractional abundance of 6-nt words; the x-axis is the location (nt) relative to the polyadenylation/cleavage site. In this case we required only one EST to establish a hypothetical poly(A) site. NUE – near upstream elements, URE – U-rich element. We used T not U here because EST is a cDNA.

RESULTS AND DISCUSSION

The poly(A) signals have been found to differ widely among yeast, animals and plants in terms of signal location and sequence content. Most plant transcripts possess several polyadenylation sites, and those are usually situated within last 3′ terminal exon in a region of 100 to 200 nt in a relatively close-packed arrangement on the order of tens nucleotides apart. 3′ Expressed Sequence Tags (EST) provide an empirical method for locating the clevage/polyadenylation site of transcripts. After applying EST clustering, alignment of cluster members clearly show possible polyadenylation sites of respective mRNA (Gautheret *et al*., 1998). Moreover transcript-genome alignments can be used to retrieve the region downstream of polyadenylation sites, which allows to study the whole set of polyadenylation signal elements. However, in plants there are only a few full sequenced genomes and at the same time huge amount of ESTs available. In our work we utilized wide spreading of multiple polyadenylation site in plants to reconstruct sequences beyond polyadenylation site and compose alignments of such extended ESTs to search cis-elements of polyadenylation signals around polyadenylation site. Fig. 2 shows results of sequence analysis around polyadenylation/cleavage site for *Oryza sativa,* which were got using genomic sequences (*a*, *b*) and EST (*c*, *d*). Nucleotide profiles of genomic (*a*) and EST (*c*) datasets are very similar, the Spearmen rank correlation coefficients are 0.97, 0.87, 0.96, 0.97 for A,G,C,T nucleotides frequencies respectively and p-value < 2.2e-16 in all cases.

To define possibility of reviling known cis-elements we employed analyses of hexamer usage (Z-score and appearance frequencies). As it is shown in Fig. 2 (*b*, *d*) we have found some patterns that corresponded to specific elements of polyadenylation signal (shown by inset on Fig. 2*b* and 2*d*) which were recently reevaluated in a model plant *Arabidopsis thaliana* (Loke *et al*., 2005). To further validate our approach we analyzed sequences around polyadenylation/cleavage site in *Triticum aestivum* (Fig. 2*e*, *f*) and were also able to find all plant specific elements of polyadenylation signal.

So proposed approach is valid enough and gives appropriate results.

REFERENCES

Brockman J.M., Singh P., Liu D., Quinlan S. *et al*. (2005) PACdb: polyA cleavage site and 3′-UTR database. *Bioinformatics,* **18**, 3691–3693.

Gautheret D., Poirot O., Lopez F. *et al*. (1998) Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. *Genome Res.,* **8**, 524–30.

Kent W.J. (2002) BLAT - The BLAST-like alignment tool. *Genome Res.,* **12**, 656–64.

Loke J.C., Stahlberg E.A., Strenski D.G. *et al*. (2005) Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures. *Plant Physiol*., **138**, 1457–1468.

Pertea G., Huang X., Liang F. *et al*. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics,* **19,** 651–652.

Schbath S. (1997) An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comp. Biol*., **4**, 189–192.

TGP (TRANSGENE PROMOTERS): A DATABASE OF BIOTECHNOLOGICALLY IMPORTANT PLANT GENE PROMOTERS

Smirnova O.G.*, Ibragimova S.S., Grigorovich D.A., Kochetov A.V.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: planta@bionet.nsc.ru

Key words: promoter, transgenesis, database, gene engineering

SUMMARY

Motivation: One of the most important problems when planning a genetic engineering experiment is to ensure the adequate pattern for transgene transcription. A database containing annotated published data on the promoters operating in plant cells with a certain specificity and activity may be used for solving this problem. Such specialized databases are yet absent.

Results: We have developed the database on promoters (TGP), collecting the information on plant promoter sequences with experimentally verified specific transcription patterns including general, tissue-, stage-, and stress-specific activities. The database was constructed on the SRS platform and consists of three cross-linked parts: gene description, promoter description, and corresponding experimental promoter sequences. TGP is aimed to provide information for experiments on transgenic plants and may be useful for either basic research in molecular biology or biotechnological experiments.

Availability: The database is available at http://wwwmgs.bionet.nsc.ru/mgs/dbases/tgp/.

INTRODUCTION

A correct planning of genetic construct design is a necessary condition for successful transgene expression. In the majority of cases, expression of a foreign gene in plant must follow a certain pattern, frequently, rather specific; for example, only in a particular tissue or at a particular developmental stage. Thus, choosing of an adequate promoter may be considered a most important stage in planning a genetic engineering experiment. In each particular case, this selection may be based on an individual systematic analysis of the relevant published experimental data; however, this is an inefficient and labor-consuming approach. A specialized database compiling the information about promoters operating in the plant cell with a certain specificity and activity may be used for solving this problem. The existing (related in the subject) information resources are incapable of solving this problem in its full value. The TRRD database is the closest to the standards of such information resource. Its section plantTRRD contains the information about plant genes and their promoters (Stepanenko *et al.*, 2000). However, the plantTRRD format is oriented mainly to transcription factor binding sites, and this information is of a limited interest for planning genetic engineering experiments. The databases PLACE (Higo *et al.*, 1999) and RARGE (Seki *et al.,* 2002). are also oriented to transcription factor binding sites. The information server AGRIS also contain the information about potential transcription factor binding sites of *Arabidopsis thaliana* (Davuluri *et al.*, 2003). The database PlantProm compiles the promoter sequences of plant genes (Shahmuradov *et al.*, 2003) but lacks the functional characterization of the promoters. Thus, the available

sources fail to provide the quantitative information about the level of promoter induction and the dependence of induction on the size of promoter fragment. Selection of nucleotide sequences in the PLACE and AGRIS databases is also a rather laborious process.

We developed the TransGene Promoter (TGP) database, containing the information about initial promoters and their deletion mutants obtained via annotation of experimental literature data. TGP comprises three constituent bases—TGP_GENE, TGP_PROMOTER, and TGP_SEQUENCE—cross-linked with one another. This provides a possibility to select promoters with the required properties including the origin, dimensions, and appropriate stress-, tissue-, and stage-specific activities for different experimental tasks. On demand, the user may retrieve the nucleotide sequence of the desired promoter as well as characteristics of the initial gene.

IMPLEMENTATION AND RESULTS

TGP is implemented on the SRS platform and consists of three constituent databases.

Database on genes (a). This database compiles the information about the genes whose promoters are offered for the transgene design. Each entry of this database contains gene and product names (GENE, PRODUCT) as well as the name of organism and its taxonomic classification (SPECIES, TAXON). The field SOURCE indicates the database wherefrom the gene nucleotide sequence was extracted (EMBL), accession number of the sequence in this database (AC), and position of the start of either transcription (ST) or translation (SR) from the entry; if it is not available, then the experimental start of transcription (STexp) from the corresponding published source is indicated. The field KEYWORD contains the name of the process wherein the gene acts and characteristics of the gene; the field DESCRIPTION details the functional activity of the gene in various organs and tissues as well as its changes during ontogenesis. The entry for a gene contains also cross-references to the TGP subdatabases PROMOTERS and SEQUENCES as well as references to the literature source wherefrom the information about this gene was extracted (Fig. 1*a*).

Database on promoters (b). This database accumulates information about functionally active promoters described in annotated scientific sources. Each entry of this database contains information grouped in 11 fields. The field LOCALIZATION specifies the promoter location relative to the start of transcription (ST, STexp) or translation (SR), reference to EMBL (AC), and location of the sequence given in the corresponding EMBL entry, for example, ST; from –1345 to +1; EMBL; U37336; from 730 to 2076.

The field REPORTER indicates the product according to which the expression pattern of a given promoter is determined (mRNA, protein, or reporter enzymatic activity). The field SPECIFICITY describes the stage of organism development, organ, tissue, cell type, and cell cycle stage of the transgenic plant where the promoter is question was studied (Fig. 1*b*). The field INDUCER contains the names of inducers that change the activity level of the promoter in question, the concentration of inducers, and their action time (Fig. 1*b*). The field COMMENT provides quantitative information about the expression and induction levels of the promoter as well as the information about specific features of the genetic construct.

Database on promoter sequences (c). This database contains nucleotide sequences of promoters annotated in the database (B). Each entry contains a full-sized promoter sequence described in the original published data, cross-reference to the promoter (PROMOTER_ID), and cross-reference to the gene (GENE_ID) whose promoter is described.

GENE_ID Pc:PR2 *a*
GENE
DATE 15.02.2006
AUTHOR Smirnova O.
GENE pr2
PRODUCT pathogenesis-related protein 2
TAXON Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta; Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; asterids; campanulids; Apiales; Apiaceae; Apioideae; apioid superclade; Apium clade; Petroselinum.
SPECIES parsley, Petroselinum crispum
SOURCE EMBL; X55736; ST: 791
KEYWORD pathogenesis, elicitor inducible
DESCRIPTION A 20- to 50-fold increase in the level of PR2 mRNA occurs within the first 3 h following addition of fungal elicitor to the cells.
PROMOTER_ID Pc:PR2_P1 Pc:PR2_P2
SEQUENCE_ID Pc:PR2_P1S Pc:PR2_P2S
REFERENCE van de Locht U., Meier I., Hahlbrock K., Somssich I.E. A 125 bp promoter fragment is sufficient for strong elicitor-mediated gene activation in parsley. EMBO J., 1990, 9, 2945-2950.
PUBMED 2390976
END

PROMOTER_ID Pc:PR2_P1 *b*
PROMOTER
GENE_ID Pc:PR2
GENE pr2
LOCALIZATION ST; from -168 to +109; EMBL; X55736; from 623 to 900
SEQUENCE_ID Pc:PR2_P1S
REPORTER GUS activity
SPECIFICITY parsley protoplasts
INDUCER elicitor (8 h)
COMMENT An 8-fold increase in GUS activity was measured in elicitor-treated versus untreated protoplasts. Deletions from position -795 down to -168 had no effect on overall GUS activity, nor on the induction factor caused by elicitor treatment.
REFERENCE van de Locht U., Meier I., Hahlbrock K., Somssich I.E. A 125 bp promoter fragment is sufficient for strong elicitor-mediated gene activation in parsley. EMBO J., 1990, 9, 2945-2950.
PUBMED 2390976
END

SEQUENCE_ID Pc:PR2_P1S *c*
PROMOTER_ID Pc:PR2_P1
GENE_ID Pc:PR2
SEQUENCE
```
  ggccaaga atgtatgttc atctttgatg tgccatgaag ttgaaattca
atagtgtgct aattgtttaa gagttgtgtc caatagggct cctgtacaat
tcaaacattg ttcaaacaag gaacctaagt tctggcaata tatataccct
ctacttcatt cattttttctt gcaccaaatt aagtttttgc tagctaatac
aagtagtatc atcattttct tgtacataat catatacaaa gtatatatta
aaattatttt tgaacgatta ttcatgggtg ctgttactac cgatgttgag
```
END

Figure 1. Examples of the TGP database entries: (*a*) gene description, (*b*) promoter description, and (*c*) full-sized promoter sequence.

The TGP database compiles the information about deletion mutants of promoters, which display different specificities and transcription activities. Annotation of these data increases essentially the success rate in selecting the desirable promoter variant. SRS tools allow for indexing the fields of databases and search by the fields via a system of adequate queries. The TGP GENE database can be searched by the following

fields: gene name (GENE), protein name (PRODUCT), species (SPECIES, TAXON), and keywords (KEYWORD). The search by the field DESCRIPTION allows the tissue-specific, organ-specific, and stage-specific genes to be displayed. This information is of special interest for the genes whose promoters were studied in cell culture, not at the level of overall organism.

The main fields for the search of the TGP PROMOTER database are SPECIFICITY and INDUCER. The field SPECIFICITY contains the description of specific expression pattern for a given promoter in particular organs and tissues and at particular developmental stages. The field INDUCER allows for searching the database by the name of inducer that influences the promoter activity. It is also possible to search TGP PROMOTER by the fields COMMENT, REFERENCE, PROMOTER_ID, GENE_ID, and SEQUENCE_ID.

Find below the typical queries to the GENE and PROMOTER databases:

(1) Find the promoters activated by a particular inducer (INDUCER, COMMENT);

(2) Find the promoters active in particular organs and tissues of transgenic plants (SPECIFICITY);

(3) Find the promoters whose activity was studied during transgenesis of particular plant species (SPECIFICITY), for example, parsley, tobacco, or Arabidopsis;

(4) Find all the genes (GENE) related to a particular process (KEYWORD, DESCRIPTION); and

(5) Find all the genes (GENE) that act in particular organs and tissues (KEYWORD, DESCRIPTION).

At present, the TGP database accumulates the information on various promoters of higher plants. Currently, it contains the data on 100 promoters, the corresponding sequences, and 30 genes. We plan to expand the content of TGP.

ACKNOWLEDGEMENTS

This work was supported by the Federal Agency for Science and Innovations (state contract No. 02.467.11.1005 of September 30, 2005), and the Program of Russian Academy of Sciences “Dynamics of Plant, Animal, and Human Gene Pools”. We thank the SD RAS Complex Integration Program (No. 5.3) and the Ministry of Industry, Science, and Technologies of the Russian Federation (grant No. 2275.2003.4) for a partial support.

REFERENCES

Davuluri R.V. *et al.* (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.

Higo K. *et al.* (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucl. Acids Res*., **27**, 297–300.

Seki M. *et al.* (2002) Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, **296**, 141–145.

Shahmuradov I.A. *et al*. (2003) PlantProm: a database of plant promoter sequences. *Nucl. Acids Res*., **31**, 114–117.

Stepanenko I. *et al.* (2000) Development of knowledge base on plant gene expression regulation. *Proceedings of the II International Conference on Bioinformatics of Genome Regulation and Structure. Vol. 1*. Novosibirsk, 185–186.

TOOL FOR AUTOMATIC DETECTION OF CO-REGULATED GENES

***Stavrovskaya E.D.*[*1, 2], *Makeev V.J.*[3], *Merkeev I.V.*[3], *Mironov A.A.*[1, 2, 3]**
[1] Institute for Information Transmission Problems, RAS, Moscow, Russia; [2] Department of Bioengineering and Bioinformatics, Moscow State University, Moscow, Russia; [3] State Scientific Center GosNIIGenetica, Moscow, Russia
[*] Corresponding author: e-mail: esta191@yandex.com

Key words: bacterial genome, regulon, clustering

SUMMARY

Motivation: To study the regulation of transcription, it is important to identify coregulated genes (regulons). One way to do that is to cluster similar potential regulatory signals found by various experimental or computational techniques, for instance, phylogenetic footprinting. This strategy does not require a priori information about gene coregulation and reports new potential components for known regulons. In addition, clustering may reveal new, currently unknown potential regulons. Such data are of particular importance for poorly annotated genomes.

Results: We have developed a computer tool for automatic detection of co-regulated genes. It implements the phylogenetic footprinting technique to find potential regulatory signals and uses the clustering procedure to identify potential regulons. The tool is intended for the analysis of sufficiently closely related bacterial genomes.

Availability: The tool was implemented in Java. The source code is available upon request.

INTRODUCTION

Predicting specific transcription regulation is arguably among the most important problems in modern molecular biology. Studies in the field employ experimental techniques, bioinformatics methods, and their combinations. A popular bioinformatics method is phylogenetic footprinting. Firstly, groups of orthologous genes are identified in a set of related genomes by protein sequence comparison. The upstream regions of orthologous genes are selected and in each group of such fragments common motifs are determined using more or less standard tools. Such a motif is considered a potential regulatory signal. This approach is of limited utility as it reports only a single site, not the complete regulon, that is, set of co-regulated gene within a genome. Indeed, a regulatory factor usually affects several genes in a genome. DNA sites bound by a particular protein are similar to each other and thus can be clustered. Thus a cluster of similar regulatory signals suggests a potential regulon.

METHODS AND ALGORITHMS

Our tool works in three main steps: identification of groups of orthologous genes, signal finding, and clustering. Dependent on a particular problem and available data, it is possible to start at any step.

To construct groups of orthologous genes we use the algorithm PHOG-BLAST (Merkeev, 2003). This algorithm was used to build the PHOG database of phylogenetic orthologous groups. PHOG-BLAST is a completely automated procedure that creates clusters of orthologous groups at each node of the taxonomy tree (PHOGs – Phylogenetic Orthologous Groups). An essential step in building this database was comparing protein complements of different species and orthologous groups of different taxa. To do it in reasonable time, PHOG-BLAST finds similarity between pairs of protein multiple alignments by converting them into "ancestral" sequences. This algorithm compares "ancestral" sequences using a special BLAST-like procedure and counts similarity scores used for finding orthologs and paralologs.

The next step is to find potential conserved regulatory signals in regions upstream of orthologous genes. To do that, we use the SignalX algorithm (Mironov *et al*., 2000). It is a greedy EM-type algorithm which uses rank statistics at different steps. The advantage of this algorithm is that it does not require that most input sequences contain a signal. SignalX efficiently reduces the number of falsely predicted sites.

The last step of our approach is clustering. We use the ClusterTree-RS algorithm (Stavrovskaya *et al*., 2006) to cluster a set of potential regulatory signals. This algorithm allows clustering signals of different lengths.

The algorithm builds the binary tree and detects clusters corresponding to its nodes. To build the tree, ClusterTree-RS compares all signal motives with each other and merges the most similar pair into a new motif. After merging, the group of signals is considered as one signal consisting of all sites of the corresponding signals. Its similarity to the remaining signals is computed as

$$D = \sum_{k=1}^{l} I_k \frac{\sum_{i=A,C,G,T} (f_1(i,k) - \overline{f}_1(k))(f_2(i,k) - \overline{f}_2(k))}{F_1 F_2},$$
$$F_j = \sqrt{\sum_{i=A,C,G,T} (f_j(i,k) - \overline{f}_j(k))^2} \qquad (1)$$

where I_k is the information content of the combined signal, $f_j(i,k)$ is the relative frequency of nucleotide i at position k of the signal j, $\overline{f}_j(k)$ is the average relative frequency in position k, $0.25\alpha\sqrt{N}$ and $\alpha\sqrt{N}$ are the pseudocounts.

Equation (1) utilizes the Pearson correlation coefficient. The coefficient assumes a maximal value when nucleotide frequencies in a given position are the same in both signals. If the correlation coefficient is used as is, the similarity between the positions with a random distribution of nucleotide frequencies will be the same as between the absolutely conserved positions. To ascribe a greater weight to conserved positions, the correlation coefficient is multiplied by the information content.

After tree-building, the algorithm considers all tree-nodes of the tree and identifies those corresponding to clusters. Each tree node corresponds to a sites set, which results from merging of two site sets corresponding to the child nodes. When the child sets of sites are similar (i.e., the corresponding nucleotide counter matrices are similar), the sets may belong to one cluster (i.e., contain the same signal). A considerable difference between the matrices suggests that the given node corresponds to the fusion of two different sets of sites, which correspond to two different signals. To establish whether the nucleotide counter matrices are similar or different, the algorithm computes the log-adds ratio:

$$R = \sum_{k=1}^{l} \log \left[\frac{(N_1 + L - 1)!(N_2 + L - 1)! \prod_{i=1}^{L} (n_1(i,k) + n_2(i,k))!}{(L-1)!(N_1 + N_2 + L - 1)! \prod_{i=1}^{L} (n_1(i,k))! \prod_{i=1}^{L} (n_2(i,k))!} \right], \quad (2)$$

where L is the alphabet size ($L = 4$), $n_j(i,k)$ is the count of nucleotide i in position k in a child node, N_j is number of sites in a child node. With this likelihood ratio, the null hypothesis suggests that signals of the right and the left child node can be obtained from the pooled matrix of the current node. A node corresponds to a cluster if the log-odds ratio is positive for this node and negative for its parent node.

All selected clusters are explored by the "noise" sites elimination procedure. Some signals identified at the previous step contain false predicted sites. Even when their fraction is low, these false sites distort the signal. It is rather difficult to eliminate such sites at an early stage, because the initial signals may few sites and there is no a priori information about the correct signal structure. However, when similar signals are clustered, there is a sufficiently large number of sites for each (now clustered) signal and the subset of statistically significant sites can be extracted. Then, all nonsignificant sites are eliminated from the cluster.

The typical runtime of the tool is as follows: identification of orthologs in a set of 30 genomes requires 5 hours; identification of candidate regulatory sites in a set of 20 350 bp fragments requires 1 minute; clustering of 30 000 motives requires 48 hours.

RESULTS AND DISCUSSION

Using the clustering procedure we have predicted some new members of known regulons in gamma proteobacteria and firmicutes. We have predicted new regulons as well. Some results are listed in Tables 1, 2 and 3.

Table 1. Analysis of the gamma-proteobacterial sample with the ClusterTree-RS algorithm

N	Regulator	Number of signals	Number of sites	Genes
1	**ArgR**	7	21	**yhcC**, argC, argA, **yjgD**, argI
2	**PurR**	10	45	purL, cvpA, codB, purM, purC[1], purE, **fold**[1], purH, purT, **yjcD**
3	**LexA**	8	35	uvrA[1], recN, lexA, ruvA, recA, **dinP**, umuD, **uvrD**
4	**MetJ**	5	24	metJ, metA, metF, metE, **yaeD**
5	**Crp**	8	36	**ycdZ,** glpA, **yiaK**, **yiaJ**, cdd, **yfiD**, **yeaA**

[1]The *E. coli* gene corresponding to the signal is indicated. However, the site was not found upstream of the *E. coli* gene, or was eliminated from the cluster as noise. The cluster contains sites found upstream of orthologous genes of related organisms.

The column Genes shows *E. coli* genes corresponding to the signals of the cluster. The genes whose regulation by the given factor is unknown according to the DPInteract database (Robison *et al.*, 1998) are set in bold.

Table 2. Analysis of the Furmicutes sample with the ClusterTree-RS algorithm

N	Regulator	Number of signals	Number of sites	Genes
1	**Cre_16 (CcpA)**	4[1]	27	**yvfK**, araE, **amyX**[2]
2	**HrcA**	4	19	**ydiL**, hrcA, groES, **htpG**[2]
3	**CtsR_aln2**	2	18	ctsR, clpE
4	**Fur**	4	34	yfiY, fhuD, feuA[2], **yqkL**

Table 3. New potential regulons identified by analysis of the EC and BS samples with the ClusterTree-RS algorithm

Set	N	Genes	Function
EC	1	nrdD	enzyme; 2'-Deoxyribonucleotide metabolism
		nrdA	enzyme; 2'-Deoxyribonucleotide metabolism
		ubiE	enzyme; Biosynthesis of cofactors, carriers:Menaquinone, ubiquinone
		proS	enzyme; Aminoacyl tRNA synthetases, tRNA modification
BS	2	pyrR	Attenuation (antitermination) of the pyrimidine operon (pyrPBCADFE) in the presence of UMP (pyrimidine biosynthesis)
		pyrP	pyrimidine biosynthesis
		pyrF[2]	pyrimidine biosynthesis
BS	3	ylpC (fapR)	Unknown
		yhfB (yhfC)	Unknown

[1] The cluster contains a signal that initially did not include a *B. subtilis* site. The signal was derived from an orthologous gene set, which included genes of Fumicutes other than *B. subtilis*. The signal included sites found upstream of the *Streptococcus pneumoniae* PN_SP2107 and *Streptococcus pyogenes* ST_malM genes.

[2] The *B. subtilis* gene corresponding to the signal is indicated. However, the site was not found upstream of the *B. subtilis* gene, or was eliminated from the cluster as noise. The cluster contains sites found upstream of orthologous genes of related organisms.

The column Genes shows *Bacillus subtilis* genes corresponding to the signals of the cluster. Genes whose regulation by the given factor is unknown according to DBTBS (Makita *et al*., 2004) are set in bold. A more interesting problem is to analyze poorly annotated genomes. We have applied our tool to genomes of alpha proteobacteria. One arising problem is that many genes in these genomes have unknown function and it is hard to say something about the common function of genes corresponding to the cluster, or to identify the transcription factor responsible for the identified signal. To solve this problem, we intend to assign functions to the groups of orthologs based on existing annotation for at least one member of the group and on COG database annotation. We will then search for GeneOnthology functions over-represented in the derived clusters.

ACKNOWLEDGEMENTS

We are grateful to M. Gelfand, O. Kalinina, and D. Rodionov for fruitful discussion. This work was supported by the Howard Hughes Medical Institute (grant No. 55005610), the Russian Academy of Sciences (program "Molecular and Cellular Biology") and the Russian Foundation for Basic Research (project No. 04-04-4938). E. Stavrovskaya acknowledges a grant from the Russian Science Support Foundation.

REFERENCES

Makita Y., Nakao M., Ogasawara N., Nakai K. (2004) DBTBS: Database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics. *Nucl. Acids Res*., **32,** 75–77.

Merkeev I.V. (2003) TOGs vs COGs: A database of supergenomes built from complete proteome sequences. *Proceedings of MCCMB'03*, 152–153.

Mironov A.A., Vinokurova N.P., Gel'falnd M.S. (2000) Software for analyzing bacterial genomes. *Mol. Biol. (Mosk.)*, **34(2)**, 253–262.

Stavrovskaya E.D., Makeev V.J., Mironov A.A. (2006) ClusterTree-RS: A binary tree algorithm identifying coregulated genes by clustering regulatory signals. *Mol. Biol. (Mosk.)*, **40(3)**, 524–532.

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

HOW SIMILAR ARE PHENOTYPICALLY IDENTICAL CELLS AT THE TRANSCRIPTIONAL LEVEL?

Subkhankulova T.*, *, Livesey F.J.
Gurdon Institute and Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge, CB2 1 QN, UK
* Corresponding author: e-mail: ts300@cam.ac.uk

Key words: single neuronal stem cell, microarray expression profile, global polyA PCR-based amplification, model distribution, statistical analysis

SUMMARY

Motivation: The expression profiling of single cells using a microarray technology requires careful interpretation of the obtained data. It is crucial to distinguish between real differences in mRNA levels, sampling effects and random technical noise.

Results: Simple mathematical models of expression data, based on sampling effect were developed and compared with real two-channel microarray data. We demonstrate that the real distribution of gene expression ratios for pairs of neuronal stem cells is much higher than predicted from sampling model.

Conclusions: These findings confirm that there is significant difference in expression levels between individual phenotypically identical stem cells.

INTRODUCTION

The improvements in microarray technology provide a tool to analyze cellular heterogeneity at the level of a single-cell gene expression profiling. Amplification of the starting mRNA population is a crucial step required to generate labeled microarray targets from limiting amounts of RNA. It has been shown that global polyadenylated PCR-based amplification technique generates reliable data from picogram amounts of RNA (Subkhankulova, Livesey, 2006). However, high variability has been reported for two-channel microarray analysis at single cell level. This variability may be caused by sampling effect (the random picking of the low abundant transcripts) and therefore depends on mRNA species abundance and the efficiency of the amplification technique. Otherwise, tested single cells can be not identical at transcriptional level even if they possess high morphological and functional similarity.

Here we provide the analysis of single cell expression data, based on estimation of efficiency of amplification technique and computational models fitted to the real distributions. We demonstrate that the real distribution of gene expression ratios for pairs of neuronal stemcells is much higher than predicted from sampling model. These findings confirm that there is significant difference in expression levels between single phenotypically identical stem cells.

MATERIAL AND METHODS

Global polyadenylated PCR amplification. Neuronal stem cells were obtained from dissections of mouse embryo neocoretex at day 11.5. Tissue was disintegrated with papain dissociation system (Worthington Biochemical Corporation) and single cells were picked by thin capillary, washed in PBS and placed in PCR tubes with cell lysis buffer following by global polyadenylated PCR amplification, firstly suggested by Hiro Matsunami (Subkhankulova, Livesey, 2006). PCR products were purified with the CyScribe GFX Purification kit (Amersham Bioscience) and labeled with Cy3/Cy5 dCTP using Klenow DNA polymerase (BD Bioscience).

Microarray hybridization. Expression microarrays containing 23232 65-mer oligonucleotides (Sigma-Genosys) were printed on Codelink slides (Amersham).

Statistical methods. All statistical analysis was conducted using the R environment and the R package 'Statistics for Microarray Analysis'. Log intensity ratios for each spot were obtained with background subtraction. Data normalization was performed using scaled loess normalization using Limma package.

MODEL

The gene expression difference between two cells obtained in microarray analysis generally may include a few components:

1. The real difference in gene expression profiles of two cells;
2. The difference caused by random picking of mRNA species from each cell RNA pool (sampling effect);
3. Technical noise arising from amplification, hybridization, washing procedures, uneven array printing, etc.

Previously we have shown that technical noise is relatively low for the microarray data obtained in the hybridizations on the oligonucleotide arrays (Subkhankulova, Livesey, 2006), therefore we ignored it in subsequent calculations. However it is impossible to distinguish the sampling effect from real difference between two single-cell samples until we know that they are completely identical. So we chouse a *single cell divided in two parts as a model* of identical samples (*model A*). The only source of diversity in expression profiles for these two parts would be uneven picking of low abundant mRNA copies. This diversity will strongly depend on number of the mRNA species (abundance) for particular gene and efficiency of the amplification technique. Then we calculated the distributions for given number of mRNA copies of particular gene from 1 to 170, assuming that if transcript's abundance is more then 170 the microarray data would reflect only technical noise:

$$p_i = (C_n^a)(C_{n-x}^{N-a}) / (C_x^N),$$

where p_i – is probability for i-th gene to be selected x times$_i$ from the mRNA pool when cell divided into two half, N – total number of transcripts in a single cell, n – number of transcripts picked from N_i, a – number of mRNA transcripts for gene i-th; x – number of transcripts for gene i-th selected from a.

To estimate the total probability distribution for transcripts with abundance from 1 to 12,000, we introduced the weight vector $W = \{w_1, w_2, \ldots w_{170}, w\}$ which represents the percentage of genes with correspondent transcript abundance, where w is the weight for genes with abundance more then 170. We fitted the model distribution to real microarray data obtained for hybridizations of half to half single cells content by optimization the weight vector. After optimization weight vector was fixed for subsequent computations.

Then we repeated the calculations for *two single cells model* (*model B*). We hypothesized that expression profiles of any two neuronal progenitor cells are completely the same. Based of estimated efficiency of amplification technique equal to 90 % and fixed weight vector (*W*) we calculated the probability function (*P*) for genes to get the given expression log ratio (*M*) using the algorithm described above. This distribution was compared with real microarray data for targets from 12 neuronal progenitor cell co-hybridized in pairs on oligonucleotide arrays.

RESULTS AND DISCUSSION

Efficiency of global polyadenylated PCR amplification. The sampling effect in generation of microarray targets depends on two factors: the absolute numbers of mRNA copies for given gene and efficiency of a few first steps of the amplification technique. The higher an efficiency of the first steps of the amplification (including cDNA synthesis, poly-adenylation, and first cycles of PCR reaction) the less mRNA transcripts are lost in fact, and the better the precision of expression profiling of target mRNA. With each cycle of PCR the efficiency of reaction becomes less important as total amount of original cDNA copies is growing and loss of 1–3 % of total number of copies is less crucial. We estimated that the first two steps (cDNA synthesis with following polyadenylation) produced 94 % of maximally expected amounts of polyadenylated ss cDNA. The PCR was as efficient as 97–98 % for each exponential cycle. Therefore, the most crucial steps of amplification of original mRNA would reproduce the original mRNA profile with approximately 90 % efficiency.

The fitting of a model distribution to real microarray data. The probability (*P*) for genes to get the given expression log ratio (*M*) was calculated based on weight vector (*W*) as described above (*model A*). The model distribution fitted the best to real M-values distribution obtained from hybridizations of half cell *vs.* half cell (Fig. 1*a*) if vector *W* corresponded to the distribution of mRNA species when very a few genes (6.5 %) demonstrate relatively high abundance (more then170 copies) and majority of genes (63 %) are represented in total mRNA pool by low numbers of transcripts (less then 50).

Comparison y expression data. We assumed that if tested single neuronal progenitor cells are absolutely identical therefore the diversity in microarray expression data will be entirely due to sampling effect, arising because of high proportion of low abundant genes and loss of transcripts during the amplification procedure. From experiments described above we estimated both these parameters: abundance of gene transcripts (*W*-vector) and the efficiency of amplification technique (90 %). Now we used these parameters to simulate the distribution of log(base2) expression ratios between two identical cells as described above (*model B*).

The distributions of real expression data are much wider then it has been predicted in our *model B*, where the diversity between two samples is due only to sampling effect (Fig. 1*b*). It means that any pair of tested single cells possesses expression difference between each other which also contribute to wide distribution of M-values. Therefore our results disapprove the hypothesis about expression identity of progenitor cells.

The variability of the transcript's levels in neuronal progenitor cells while they posses high morphological and functional similarity may be a result of stochastic fluctuations intrinsic normal alive cells (Levsky, Singer, 2003).

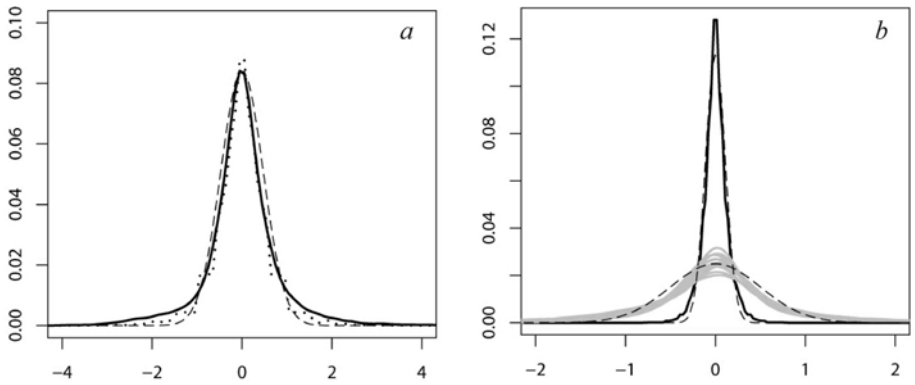


Figure 1. The model distribution of M-values (log (base2) expression ratios) fitted to real microarray data (*a*). Dashed line – model distribution based on optimized weight vector (model *a*); black line – real distribution of M-values, obtained for half *vs*. half of single cell; pointed line – theoretical Gaussian distribution (sd = 0.42). The distributions of log(base2) expression ratios (M) for pairs of real cells are higher then predicted for two identical samples (*b*). Black solid line – predicted distribution of M-values for two identical samples; gray lines – real distributions of expression ratios for pairs of 12 neuronal progenitor cells. Dashed lines – theoretical Gaussian distributions with sd = 0.5 (approximation of average microarray expression data) and sd = 0.11 (approximation of the model distribution).

CONCLUSIONS

1. We developed statistical models that can be used to validate of a single cell microarray expression data.
2. Our results show that both sampling effects and different expression levels contribute to the wide distribution of log(base2) ratios obtained for two-channel microarray analysis of pnenotypically similar cells.
3. Neuronal stem cells demonstrate high heterogeneity which possibly is a result of stochastic fluctuations in mRNA transcript levels intrinsic in cycling cell.

REFERENCES

Subkhankulova T., Livesey F.J. (2006) Comparative evaluation of linear and exponential amplification techniques for expression profiling at the single cell level. *Genome Biol*., **7**(3).

Levsky J.M, Singer R.H. (2003) Gene expression and the myth of the average cell. *Trends in Cell Biology*, **13**(1), 4–6.

THE MODIFIED FUZZY C-MEANS METHOD FOR CLUSTERING OF MICROARRAY DATA

***Taraskina A.S.*[*1, 3], *Cheremushkin E.S.*[2, 3]**
[1] Mechanics and Mathematics Department, Novosibirsk State University, Novosibirsk, 630090, Russia;
[2] Biorainbow Group, Novosibirsk, Russia; [3] Ershov Institute of Informatics Systems, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: anna@biorainbow.com

Key words: fuzzy clustering, microarray analysis, genetic algorithm

SUMMARY

Motivation: Microarray experiments provide large amount of data, thus development of appropriate methods and tools of analysis is rather important. Clustering as one of such methods allows identifying biologically relevant groups of genes.

Results: We developed an algorithm of the fuzzy c-means family, designed for clustering of microarray data and distance matrices, with genetic algorithm as optimization.

Availability: The program, which implements the developed method and some additional features, is available at http://biorainbow.com/fuzzyclustering/

INTRODUCTION

DNA microarrays are used to monitor gene expression in many areas of biomedical research. To analyze the increasing amount of data produced by this technology, clustering has become inevitable, see (Golub, 1999).

Clustering methods are divided into hierarchical and partitional ones. Hierarchical algorithms associated with dendrogram construction are good for a small number of objects and are not suitable for a large volume of data due to laboriousness of the agglomerative algorithm. In partitional algorithms, the data are immediately divided into several clusters, whose number is estimated depending on conditions. Then the elements are transferred between clusters to optimize a certain criterion, for example, to minimize variation within clusters. Partitional clustering methods assign each gene to a single cluster, but information about the influence of a given gene for the overall shape of clusters also makes sense. In this case fuzzy clustering is more suitable.

This work was aimed at the development and implementation of a clustering algorithm based on the fuzzy c-means in association with genetic algorithm, see (Hall *et al*., 1999) to provide a close to optimal solution to the problem of clustering of the given microarray dataset.

METHODS AND ALGORITHMS

Fuzzy c-means algorithm. (http://matlab.exponenta.ru/fuzzylogic/book1/index.php) Input information for clustering is the matrix of observations ($l \times n$ matrix) $X = [x_{ij}]$,

where l is the number of objects, n is the number of characteristics (observations) of each object.

The task of clustering is to partition an array of objects into groups (clusters) of objects that are "similar" to each other. In the n- dimensional metric space of characteristics, let us consider the distance between two objects as the measure of their "similarity".

The present work uses the fuzzy clustering method allowing each object to belong to several or all the clusters simultaneously with different degrees. The number of clusters c is considered *a priori* known.

The cluster structure is specified by the membership matrix ($c \times l$ matrix) $M = [m_{ij}]$, $m_{ij} \in [0,1]$ is the membership value of the *j*-th element to the i-th cluster, satisfying the following conditions: 1) $\sum_{i=1}^{c} m_{ij} = 1, j = \overline{1,l}$ and 2) $0 < \sum_{j=1}^{l} m_{ij} < l, i = \overline{1,c}$.

To assess the quality of partitioning, the dispersion criterion is used. It shows the sum of distances from objects to the centers of clusters with the corresponding membership values: $J = \sum_{i=1}^{c} \sum_{j=1}^{l} (m_{ij})^w d(v_i, x_j)$, where $d(v_i, x_j)$ is the Euclidian distance between the object $x_j = (x_{j1}, x_{j2}, ..., x_{jn}) \frac{\pi}{3}$ and the center of cluster $v_i = (v_{i1}, v_{i2}, ..., v_{in})$, $w \in (1, \infty)$ is the exponential weight determining the fuzziness of clusters, $V = [v_{ij}]$ is the $c \times n$ matrix of coordinates of the centers of clusters whose elements are calculated according to the formula $v_{ik} = \frac{\sum_{j=1}^{l} (m_{ij})^w x_{jk}}{\sum_{j=1}^{l} (m_{ij})^w}, k = \overline{1,n}$. (1)

The task is to find matrix *M*, which minimizes criterion *J*. To do this, the fuzzy *c*-means algorithm based on the method of Lagrangian multipliers is used. It allows us to find the local optimum, that's why different results can be obtained for different initiation processes.

At the first step, the membership matrix *M* satisfying the conditions above is generated in a random way. Then the iteration process for the calculation of the clusters centers and the recalculation of the elements of the membership matrix values is initiated:

$$m_{ij} = \left((d_{ij})^{\frac{2}{w-1}} \sum_{k=1}^{c} \frac{1}{(d_{kj})^{\frac{2}{w-1}}} \right)^{-1} \text{ at } d_{ij} > 0 \text{ and } m_{kj} = \begin{cases} 1, k = i \\ 0, k \neq i \end{cases} \text{ at } d_{ij} = 0,$$

where $d_{ij} = d(v_i, x_j)$ for $i = \overline{1,c}, j = \overline{1,l}$.

The calculations are continued until the change in matrix $\left\| M - M^* \right\|^2$, where M^* is the matrix at the previous iteration, becomes smaller than the preset stopping parameter ε.

Let us consider the selection of the exponential weight value. The larger is this value the fuzzier is the matrix of membership, and at $w \to \infty$ $m_{ij} = 1/c$, i.e. all the objects are distributed among all clusters uniformly. Usually $w = 2$ is set, but it was found out that this value is not suitable for data produced with microarrays. For the calculation of a more suitable value our program uses experimentally determined formulas, see (Dembele, Kastner, 2003). m_{ij} values depend on the distances between the elements and the centers

of clusters. The centers of clusters are close to some elements (genes), that's why it can be supposed that there exists interrelation between the results of fuzzy clustering and the coefficient of variation cv for the array $Y_w = \left\{ d(x_i, x_j) \right\}^{\frac{2}{w-1}}, i \neq j = \overline{1,l}$ where $cv = \sigma(Y_w) / \overline{Y_w}$. According to the experimental results, the equation $cv(Y_w) \approx 0.03n$, where n is the dimension of data, was proposed for determining boundary value w_{ub}.

Finally, the parameter value is chosen as follows: $w = 1 + w_0, w_0 = \begin{cases} 1, w_{ub} \geq 10 \\ w_{ub} / 10, w_{ub} < 10 \end{cases}$.

Genetic algorithm. (Goldberg, 1989). The local minimum obtained with the fuzzy *c*-means algorithm often differs from the global minimum. The search for the global minimum of functional *J* can't be realized due to a large volume of calculations, but there exist algorithms obtaining a solution close to the global minimum. We used a genetic algorithm based on genetic processes of biological organisms: biological populations develop during several generations obeying natural selection laws and according to the principle "only the fittest survives". Usually, GAs give good results for parametric functions optimization problems, and it is a problem of this type that we are solving. However, like other methods of evolutional calculations, they do not guarantee finding the global solution during polynomial time. GAs do not guarantee that the global solution will be ever found, but they are good for seeking a "sufficiently good" solution to a problem within a "sufficiently short" time.

Silhouette. The silhouette value can be used to assess the quality of clustering, see (Dembele, Kastner, 2003). Suppose gene x_i is in cluster C_r. At fuzzy clustering, the number of the cluster is determined by the maximal value of the degree of membership. The following values are calculated $a(x_i) = \frac{1}{|C_r|} \sum_{x_j \in C_r} d(x_i, x_j)$ and $b(x_i) = \min \left\{ \frac{1}{|C_s|} \sum_{x_j \in C_s} d(x_i, x_j), r \neq s = \overline{1,c} \right\}$. The gene silhouette is determined as $s(x_i) = \frac{a(x_i) - b(x_i)}{\max(a(x_i), b(x_i))}$. The silhouette value is within the interval $[-1;1]$; if it is negative, the gene is considered poorly clustered.

IMPLEMENTATION AND RESULTS

Microarrays. The data produced as a result of experiments with microarrays can be presented in a form of matrix, where the lines will contain different genes, and the columns will contain their expression levels in different experiments. The Euclidian distance is taken as the distance between the genes. The coordinates of the centers of clusters are determined according to the formulas (1). If the data are normalized (the zero mean level of expression for each gene and the single mean square deviation), then clustering gives groups of genes with similar expression profiles. Otherwise genes with close expression values fall into the same cluster during all the experiments.

The program also deals with somehow produced matrices of distances between objects and similarity matrices.

The results of clustering are partially displayed in the program window as a list of elements for clusters with the degree of membership higher than the threshold value. The

stored file with the results contains algorithm parameters, list of the genes for clusters with the degree of membership higher than 1/c, matrix of memberships, coordinates of the centers of clusters, silhouette values if calculated by the user

A test example. The algorithm functioning was tested on data sets produced in experiments on cell cycle investigation, which can be found at the site http://genome-www.stanford.edu/Human-CellCycle/Hela/.

Normalized expression values for genes participating in the cell cycle regulation, which were measured with 1-hour periodicity were taken for clustering. We performed the partitioning of genes into 5 clusters according to the number of cell cycle stages. The corresponding stage and, therefore, the cluster were predicted for each gene using the algorithm of hierarchical clustering, see (Whitfield *et al*., 2002). Suppose that such predicted distribution is exact; then the ratio of the maximal number of genes in a stage falling in one cluster to the total number of genes of this stage characterizes the accuracy of clustering with our algorithm. For some stages obtained results conform with sufficiently high accuracy (above 80 %), for some stages they do not. One of the reasons can be the similarity of expression profiles of genes of close stages (G2, G2/M). It is also quite probable that preliminary distribution with the hierarchic algorithm differs from the true one

CONCLUSION

Fuzzy clustering with the c-means method presents a convenient approach to isolation of genes tightly associated with preset clusters. Using it in a combination with the genetic algorithm, one can find a close to optimal solution to the problem of clustering.

We wrote a program for clustering, in which the above methods are realized. In addition, the program provides for the possibility of automatic identification of the value of the fuzziness parameter of clusters, which is suitable for a concrete data type, and the evaluation of the quality of clustering.

Besides, our program can be used to divide objects into groups knowing only the pair distances between them, without thinking about the coordinate presentation of these objects.

The program is realized in Microsoft Visual Studio environment and is available at http://biorainbow.com/fuzzyclustering/.

REFERENCES

Dembele D., Kastner P. (2003) C-means method for clustering microarray data. *Bioinformatics*, **19**(8), 973–980.

Goldberg D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, Mass.

Golub T.R. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Hall L.O., Ozyurt I.B., Bezdek J.C. (1999) Clustering with a genetically optimized approach. *IEEE Transactions on Evolutionary Computation,* **3**(2), 103–112.

Whitfield M.L., Sherlock G. *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.

MULTIPLE COLLAPSE CLUSTERING

***Tatarinova T.*[*1], *Schumitzky A.*[2]**
[1] Ceres, Inc. Thousand Oaks, California, USA; [2] University of Southern California, Los Angeles, USA
[*] Corresponding author: e-mail: ttatarinova@ceres-inc.com

Key words: Gene Expression Clustering, time series, Kullback-Leibler distance

SUMMARY

In this manuscript, we present our Multiple Collapse Clustering (MCC) method for treatment of data-rich problems.

Motivation: MCC is not limited to clustering of genes by similarity of their expression pattern: we suggest to compute parameters of piecewise continuous functions that approximates each gene. Our method is based on clustering of parameters of such curves.

Results: We have developed a new method to analyze gene expression time series data. As a result of our clustering procedure for each cluster we obtain a smooth centroid curve and a set of curve mean parameters and standard deviations. On a test set MCC performed better compared to the K-means clustering.

INTRODUCTION

A number of great methods were developed for clustering of gene expression data. The choice of the method depends mainly on the data representation. Gene expression data can be either in the form of the vector of measured intensities (or ratios) or in the relational form (i.e., as correlation coefficients between pairs of genes (Yang *et al*., 2000). Clustering algorithms are divided into supervised, when a set of reference clusters is known, unsupervised, and hybrid. An example of the supervised clustering algorithm is a fashionable Support Vector Machines method, which can learn the decision boundaries between data classes. Popular unsupervised clustering algorithms include Self-Organizing Maps, K-means and hierarchical clustering.

We developed a new method to analyze gene expression time series data. We assumed that the underlying biological process responsible for the change of mRNA levels in a cell can be described by a piecewise continuous function. We propose to approximate parameters of this function using multivariate normal distribution. Our method finds parameters of such function for individual genes and then genes are clustered based on the values of these parameters rather than observed expression. As a result of the clustering procedure, we obtain a smooth centroid curve and a set of curve mean parameters and standard deviations for each cluster. We believe that this approach better reflects biological continuity of cell processes.

METHODS AND ALGORITHMS

We would like to suggest a new method for clustering of data-rich time-series observations. Our approach assumes that every cluster can be described by a smooth centroid curve. If N genes can be grouped into K groups by their expression profile, then

all genes that belong to a group k = 1, …, K have similar values of the "trajectory" parameters θ, where θ is an n-dimensional vector. Our proposed Multiple Collapse Clustering (MCC) method is based on the extension of the idea of Sahu and Cheng (Sahu, Cheng, 2002) to use the weighted Kullback-Leibler distance to find the optimal number of mixture components. The difference between the method of Sahu and our method is that we suggest collapsing multiple components following only one run of the Gibbs sampler.

At the first step of MCC we assume that all genes are allocated to separate clusters. For every gene I = 1, ..., N, values of θ_i can be found by analyzing the following model in WinBUGS (Spiegelhalter *et al*., 2003):

$$\begin{aligned} &p(y_{ij} \mid \theta_i, t_j, \tau) = N(y_{ij} \mid f(t_j, \theta_i), \tau^{-1}) \\ &p(\theta_i \mid \mu_i, \Sigma_i) = N(\theta_i \mid \mu_i, \Sigma_i) \\ &p(\tau \mid \nu_0, \tau_0) = \Gamma(\tau \mid \frac{\nu_0}{2}, \frac{\nu_0 \tau_0}{2}) \\ &p(\Sigma_i^{-1} \mid R, \rho) = W(\Sigma_i^{-1} \mid (R\rho)^{-1}, \rho) \end{aligned} \tag{1}$$

where j = 1, ..., T and $f(t_j,\theta_i)$ is some nonlinear function. For simplicity we assume independence of the parameter vector components. This assumption greatly increases the speed of computation, but it is not necessary from the theoretical point of view. Note, that Equation 1 is not a mixture model. The fitting of Equation 1 requires a data-rich situation (T >> dim(θ_i). At convergence, the posterior means of μ_1, ..., μ_N and Σ_1, ..., Σ_N are obtained, and denoted by the same symbols. At the second step we construct a "Pseudo-mixture model" using parameter values estimated in the first step: $F^{(N)} = \sum_{i=1}^{N} \frac{1}{N} N(\theta \mid \mu_i, \Sigma_i)$. The traditional Kullback-Leibler distance $d_{k,k'} = \int f^{(N)} \log f^{(N)} \Big/ f_{k,k'}^{(N-1)}$, where k, k' =1, …, N between the original and the collapsed versions is then analytically computed for P^N_2 pairs of genes. For each gene k the "nearest" gene k' is found, and if the distance $d_{k,k'}$ is below a certain threshold, the genes are assigned to the same cluster. At the next step, the algorithm starts with $\frac{N}{2} \le K^{(1)} \le N$ clusters, some of them are singletons and some contain two members. A new $K^{(1)}$ by $K^{(1)}$ matrix of distances is computed between original and collapsed versions for all clusters, nearest neighbors are identified for each cluster, and clusters are merged if the distances are below the threshold. The process is repeated until there are no more clusters to be merged.

IMPLEMENTATION AND RESULTS

To check the validity and limitations of the MCC approach we simulated T = 15 time points for 100 genes evenly partitioned into K=10 clusters. We have chosen the trajectory $f(t_j,\theta_i)$ to be $f(t_j, \theta_i) = \theta_i^{(1)} + \theta_i^{(2)}(t_j - \theta_i^{(3)})\eta(t_j - \theta_i^{(3)}) \exp(-\theta_i^{(4)}(t_j - \theta_i^{(3)}))$, where random vector θ_i has multivariate normal distribution and η() is a step function. Generated values of θ_i are shown in the Table 1 below.

Table 1. Parameters of cluster mean curves

Cluster	Θ^1	Θ^2	Θ^3	Θ^4
1	1	1.086	4.394	0.6011
2	1.101	1.136	4.257	0.3112
3	7.072	0.6165	1.528	0.0598
4	3.014	0.3739	4.328	0.3325
5	0.5453	0.7643	8.342	0.3184
6	4.528	0.4004	13.79	0.2319
7	0.7576	0.3784	7.815	0.2938
8	0.249	2.174	10.31	0.2991
9	0.4947	1.224	4.883	0.1531
10	2.825	0.9269	6.244	0.2999

We ran 200,000 iterations on WinBUGS discarding the first 50,000 iterations as a burn in. The simulation took 30 minutes on a Windows NT PC (processor Intel Pentium 2.2GHz, 1GB RAM). Post processing MCC written in C++ took approximately one minute to complete. It has recovered the correct number of clusters and produced correct values of cluster parameters (Table 2). Only 7 subjects were assigned to wrong clusters. In comparison, K-means clustering in implementation of Tseng (Tseng, 2005) "misplaced" 26 subjects.

Table 2. Simulated parameters of cluster mean curves

Cluster	Θ^1	Θ^2	Θ^3	Θ^4
1	1.00631	0.98939	4.3556	0.56018
2	1.0291	0.98049	4.252	0.3118
3	6.9405	0.66686	1.6311	0.05989
4	2.89043	0.540829	4.3271	0.34505
5	0.65213	0.593069	8.4598	0.31822
6	4.543	0.65446	13.714	0.2594
7	1.09162	0.5234	7.8505	0.30562
8	0.3718	2.0025	10.262	0.28677
9	0.5308	1.23579	4.8775	0.15281
10	2.91466	0.673742	6.1841	0.316658

DISCUSSION

The proposed method may be used as a method of clustering of time series data with an unknown number of clusters, providing not only the cluster membership as its output, but also a mathematical model of gene behavior. Although the total number of genes on a genome is computationally prohibitive to be analyzed by WinBUGS, the pre-processing step can reduce the problem to a manageable size by eliminating those genes that do not show differential expression during the experiment of interest.

Multiple Collapse Clustering method performed better as compared to K-means clustering method on the simulated dataset analysis, even though K-means clustering was supplied the correct number of parameters. We plan to work on improving this result by utilizing properties of model parameter distributions found by the Gibbs sampler. The disadvantage of K-means method is that it requires *a priori* knowledge of the ultimate number of clusters and does not utilize the parameter distribution information for each gene.

One of the biggest problems with the Multiple Collapse Clustering is its strong dependence on the WinBUGS version of the Gibbs Sampler that makes them inherit all WinBUGS limitations. In order to make it into a self-sufficient tool, it is desirable to implement the Gibbs sampler inside Multiple Collapse Clustering. We plan to develop a theoretical approach to find cut-off parameters for Multiple Collapse Clustering, and compare various clustering strategies based on the weighted Kullback-Leibler distance.

The Greedy nature of the nearest neighbor method may sometimes link members from two close clusters. The advantage of the nearest neighbor clustering method is its speed. However, for more refined results, we plan to look for slower and more sophisticated methods of clustering in the future.

REFERENCES

Sahu S.K., Cheng R. (2002) A Fast Distance Based Approach for Determining the Number of Components in Mixtures. *Technical Report at University of Southampton*.

Spiegelhalter D. *et al.* (2003) WinBUGS User Manual, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, UK, Version 1.4.

Tseng G.C., Wing H. Wong (2005) Tight clustering: a resampling-based approach for identifying stable and tight patterns in data. *Biometrics*, **61**, 10–16.

TOWARDS THE IDENTIFICATION OF ANTISENSE RNAS WITHIN GENES OF TRANSCRIPTION REGULATORS

Tutukina M.N., Masulis I.S., Ozoline O.N.*
Institute of Cell Biophysics, RAS, Pushchino, Moscow region, Russia
* Corresponding author: e-mail: ozoline@icb.psn.ru

Key words: *Escherichia coli*, antisense transcription, regulatory RNAs

SUMMARY

Motivation: Almost one hundred of small regulatory RNAs (sRNAs) have been discovered in bacteria. Most of them are encoded *in trans* to regulated genes, while the set of known antisense transcripts, generated from within coding sequences (aRNAs) include only few species. Currently it is not clear how accurately this difference reflects the real situation. Thus, most methods used to reveal regulatory RNAs in genome-wide scale were based on searching for evolutionary conserved sequences or were purposefully attuned to intergenic regions. That prohibited identification of novel transcripts within coding sequences. To our knowledge a total of 1493 non-overlapping genes for untranslated RNAs have been suggested in "empty" genomic regions. To estimate the significance of antisense transcription we used pattern recognition software PlatProm, capable of predicting promoters independently on their location. More than a thousand of promoter-like signals for antisense transcription have been found. Direct experimental verification of their activity is required to estimate a reliability of these predictions.

Results: Here we analyze functional attribution of genes, possessing putative promoters for antisense transcription, characterize the distribution of promoter-like signals in the genetic locus containing gene *hns* and provide experimental evidence that RNA polymerase *in vitro* forms transcriptionally competent complexes with its internal promoter for antisense transcription. We also compare free energies of folding for known sRNAs with those of aRNAs, predicted in genes of transcription factors, and conclude that aRNA which may be expressed from *hns* has stability typical for other aRNAs.

Availability: Genomic coordinates of predicted transcription start points for antisense transcription are available by request (ozoline@icb.psn.ru).

INTRODUCTION

Small regulatory RNAs act by multiple mechanisms utilizing RNA-RNA base pairing and regulate translation of mRNAs as well as their processing and stability. Unlike the plasmid, bacteriophage or transposon aRNAs, which are transcribed from the opposite strand of their target genes, most bacterial sRNAs are expressed from their own genetic loci (reviewed in: Ozoline, Deev, 2006). The possibility of antisense transcription from bacterial genes was however testified by the data of microarray analysis (Selinger *et al*., 2000); directional cloning of short RNA species (Vogel *et al*., 2003; Kawano *et al*., 2005a) and a promoter cloning technique (Kawano *et al*., 2005a). The later experimental approach gave the largest contribution to the set of experimentally verified antisense RNAs. Since most aRNAs should be produced from their own promoters we also exploited this feature,

trying to predict promoters for antisense transcription by computational search (Brok-Volchanski *et al*., 2005). In this study it has been found that significant fraction of genes, containing such promoters, encode regulatory proteins. That is why we selected promoters for antisense transcription predicted within gene *hns* as the first candidates for experimental verification. After it has been observed that RNA polymerase forms transcriptionally competent complexes with at least one selected promoter, we compared folding propensity of putative aRNAs with that of known untranslated RNAs so as to evaluate, how typical for aRNAs are stable secondary structures.

METHODS AND ALGORITHMS

Genome-wide searching for potential promoters has been done by promoter searching software PlatProm (Brok-Volchanski *et al*., 2005).

Ability of RNA polymerase of interacting with predicted promoters was tested by gel-retardation assay and potassium permanganate footprinting. Transcription complexes were formed at 35 °C in standard buffer, containing 50 mM Tris-HCl (pH 8.0), 0.1 mM EDTA, 0.1 mM DTT, 10 mM $MgCl_2$, 50 mM NaCl, BSA (5 mg/ml), 0.2 pm of ^{32}P-labeled DNA and 1 pm of RNA polymerase. Interaction was allowed during 30 min. In gel-shift experiments a 20 mkg/ml of heparin was added before loading the sample on 5 % polyacrylamide gel pre-warmed to 35 °C. Gels were run at constant temperature until bromphenol blue migrated to the bottom of the gel. Potassium permanganate footprinting was performed as described (Zaychikov *et al*., 1997). The products of digestion were separated in 8 % polyacrylamide gel in the presence of 8M urea. Bands were visualized by radioautography.

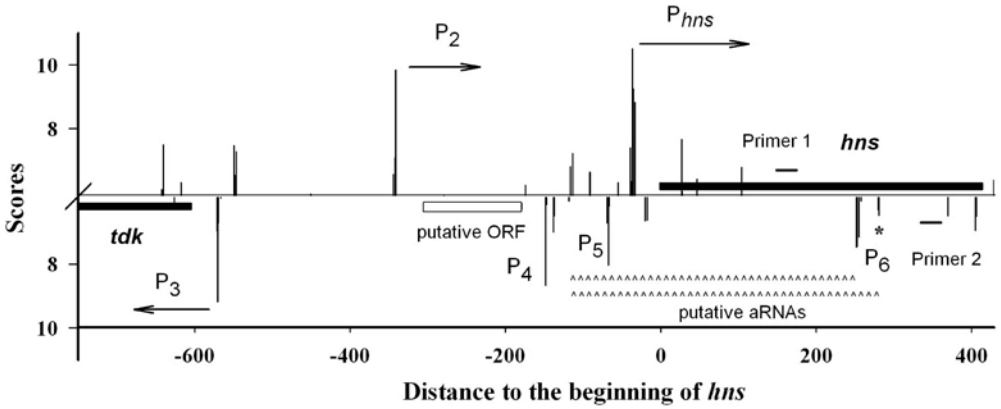


Figure 1. Schematic representation of the genetic locus, containing *hns*. Solid black lines drown above or below X axis show coordinates of *hns* and *tdk*. Bars represent promoters predicted on both strands. Only signals with $p < 0.0001$ are shown. Open rectangle indicates location of putative ORF. Two caret lines show putative RNA products, which may be synthesized between promoters P_6 and P_6^* and the first ρ-independent terminator.

Transcription terminators were searched downstream from predicted promoters on the basis of next criteria: ≥ 5 bp G/C-rich stem, 3–8 bases loop, free energy of folding less than -7 kcal/mol, at least 4 uridine residues downstream from the stem (Argaman *et al*., 2001). Folding propensities of known sRNAs and potential aRNAs were estimated by means of RNA Structure algorithm supplied with thermodynamic scoring system (http://rna.chem.rochester.edu). The set of known regulatory RNAs was taken from compilations published by Kawano *et al*. (2005b) and Ozoline, Deev (2006). The search for alternative ORFs has been done using ORF Finder (www.ncbi.nlm.nih.gov).

RESULTS AND DISCUSSION

The *E. coli* genome scanning by PlatProm revealed 1192 genes, which have internal promoter-like sites for antisense transcription. Most of them are located near the gene ends thus asuming the synthesis of long aRNAs. There are also 126 promoters, which are found less than 50 bp far from the 3′-end of genes. All together they comprise 1318 promoter-like regions, which may account for antisense transcription. Part of them may be required to transcribe upstream genes (~18 %) if they have similar orientation with promoters, or express new genes from within intergenic loci. Since in both cases putative transcripts still may function as aRNAs, such promoters were not eliminated from further analysis. Only 4.8 % of predicted promoters are found within 50 bp distance from the beginning of genes, where transcription start points for aRNAs affecting translation initiation are traditionally expected.

The whole set of genes, which may be subjected to antisense regulation was classified in respect to functions of encoded proteins. Besides enzyme-coding genes, which are dominant in any genome, and genes with unknown functions, it includes many species encoding membrane, transport and regulatory proteins, as well as proteins participating in DNA and RNA synthesis and processing. Two groups of genes, encoding proteins for RNA processing and DNA binding transcription regulators are over presented in this set. Thus, 33.9 and 33.8 % of genes ascribed to these two categories contain potential promoters for aRNAs, while the percentage of enzyme-encoding genes is lower (29.8 %). Assuming a possibility of feedback regulation in the expression of transcription factors, this observation leads us to select such genes for experimental verification. Fig. 1 shows genomic organization near *hns*. It encodes a DNA-binding protein affecting expression of many genes and is transcribed from the promoter P_{hns}, exactly predicted by PlatProm. There are many other promoter-like signals: P_2 may intensify expression of *hns*, P_3 most probably controls mRNA synthesis of *tdk* (lies on the opposite strand), P_4 and P_5 may be required to transcribe putative ORF found by ORF Finder. The synthesis of aRNA may be initiated from promoters P6 and P_6*.

The scores of two P_6 promoter-like signals are not high, however we found that RNA polymerase is capable of interacting with corresponding DNA fragment (Fig. 2*a*). Although the binding constant is low (much DNA remains free), bound RNA polymerase forms transcriptionally competent complexes, since potassium permanganate footprinting revealed the presence of unpaired thymines (Fig. 2*b*). Positions of reactive thymines correspond to the weaker promoter P_6^* (genomic coordinates 1291865), which, however, does not exclude a possibility that complexes were also formed near both predicted start points, since in the region of expected DNA melting near the stronger promoter-like site there is only a single thymine residue (expected length of the fragment 116 nt), which in this particular promoter might be protected by the enzyme. In any case, RNA polymerase forms transcriptionally competent complex with internal promoter predicted for antisense transcription.

The first ρ-independent terminator was found 388 bp downstream from P_6^* (or 360 bp from P_6). This is within the range of lengths, typical for bacterial sRNAs. Both aRNA products may form secondary structures with free energy of folding -109.4 and -89,2 kcal/M, respectively. Since absolute values are not informative, we compared the stability of these RNAs with folding propensity of known sRNAs. RNA Structure algorithm was used to estimate free energy of folding for a total of 94 sRNAs, including 11 species of short (32–80 nt) aRNAs (Fig. 3). This comparison clearly indicates that stability of both aRNAs predicted within *hns* is slightly lower than expected for sRNAs of the same length. However structural features of aRNAs are not characterized so far. It is not clear how important is their capacity to form stable secondary structures and how large are typical values of their free energy of folding. Short length of known aRNAs does not allow direct comparison, while small number excludes adequate extrapolation. We, therefore try to compare stability of these products with that of another aRNAs, predicted within genes of transcription regulators (open rectangles in Fig. 3). Free energy of folding was calculated for putative aRNAs synthesized between promoters predicted within genes of transcription regulators and the first ρ-independent terminator. Sequence motifs, suiting to the formal criteria of

transcription terminators were found within 1000 bp distance downstream of 64 out of 77 promoters. The remaining 13 aRNAs may be longer or their synthesis may be stopped at ρ-dependent terminators. Fig. 3 demonstrates that RNA product, transcribed from P_6^*, has folding propensity typical for other predicted aRNAs.

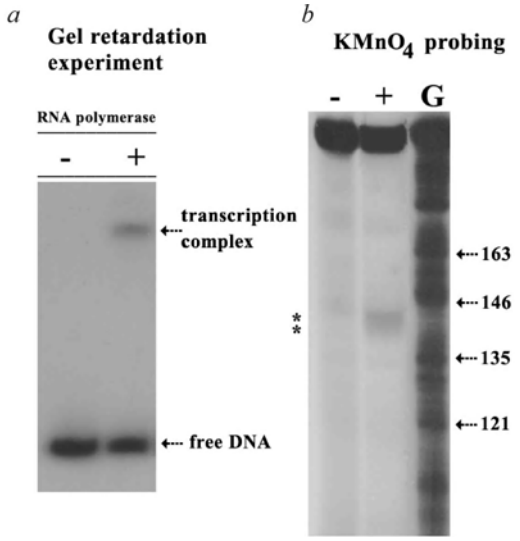


Figure 2. Experiments verifying RNA polymerase binding activity by means of gel shift assays (*a*) and $KMnO_4$ footprinting (*b*). PCR amplified DNA fragment (^{32}P-labeled primer 1 and primer 2) was used as a template. Complexes were formed under standard reaction conditions as described in Methods and Algorithms and used either for gel-retardation experiment (*a*) or for footprinting (*b*). RNA polymerase:promoter ratio was 1:5(M:M). Marks "–" and "+" denote samples, containing free DNA and DNA-protein complexes, respectively. G-specific ladder of the same DNA fragment was used to calibrate the gel. Ciphers on the right reflect sizes of indicated fragments. Asterisks on the left indicate fragments appeared due to the reactivity of unpaired thymines.

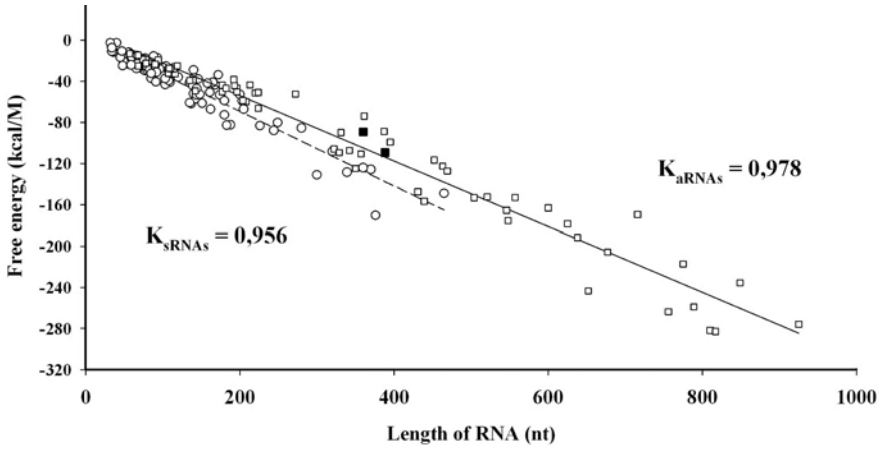


Figure 3. Correlation between free energy of folding and the size of RNA molecules. Open circles correspond to 94 known untranslated RNAs, while open rectangles correspond to putative aRNAs within genes of transcription regulators. Black rectangles show aRNAs, which may be transcribed from P_6 and P_6*.

The free energy of folding strongly correlates with the length of RNA molecule. Significant deviations from expected values take place if nucleotide sequences permit

forming of long hairpins or are depleted in inverted repeats, which are required for base pairing. For the set of predicted aRNAs the value of correlation coefficient (**K**) is 0.978, assuming strong dependence on size. For the whole set of known regulatory RNAs **K** is smaller (0.956). This value further decreases (0.954) if 11 species of known aRNAs are eliminated from this set, assuming some specificity in their structural organization. Both these observations, as well as the fact that the first order regression lines do not overlap, indicate smaller stability of analyzed aRNAs comparing to another species of sRNAs.

Thus, the first attempt to use computational approach to find genes subjected to antisense regulation revealed a large number of internal promoters, able to control production of aRNAs. The sets of predicted genetic loci encoding regulatory RNAs transcribed *in cis* and *in trans* appeared to be comparable in size. Promoter for antisense transcription (P_6^*), predicted within gene encoding global transcription regulator Hns and analyzed in this study, exhibited features of classical bacterial promoter. An expected aRNA synthesized from P_6^* has free energy of folding typical for aRNAs of the same length. Our data also indicate that the stability of aRNAs transcribed from coding sequences may be lower than that of sRNAs encoded by independent genes.

ACKNOWLEDGEMENTS

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

REFERENCES

Argaman L., Hershberg R., Vogel J., Bejerano G., Wagner E.G.H., Margalit H., Altuvia S. (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol*., **11**, 941–950.

Brok-Volchanski A.S., Masulis I.S., Shavkunov K.S., Lukyanov V.I., Purtov Yu.A., Kostyanicina E.G., Deev A.A., Ozoline O.N. (2005) Predicting sRNA genes in the genome of *E. coli* by the promoter-search algorithm PlatProm. In Kolchanov N., Hofestaedt R., Milanesi L., (eds), *Bioinformatics of Genome Regulation and Structure II*. Springer, pp. 11–20.

Kawano M., Reynolds A.A., Miranda-Rios J., Storz G. (2005a) Detection of 5′- and 3′-UTR-derived small RNAs and cis-encoded antisense RNAs in *E. coli*. *Nucl. Acids Res*., **33**, 1040–1050.

Kawano M., Storz G., Rao B.S.,Rosner J.L., Martin R.G. (2005b) Detection of low-level promoter activity within open reading frame sequences of *E.coli*. Nucl. Acids Res., **33,** 6268–6276.

Ozoline O.N., Deev A.A. (2006) Predicting antisense RNAs in the genomes of *E.coli* and *S.typhimurium* using promoter-search algorithm PlatProm. J. Bioinf. Comp. Biol., **4**(2) (in press).

Selinger D.W., Cheung K.J., Mei R., Johansson E.M., Richmond C.S., Blattner F.R., Lockhart D.J., Church G.M. (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome arrays. *Nature Biotechnol*., **18**, 1262–1268.

Vogel J., Bartels V., Tang T.H., Churakov G., Slagter-Jager J.G., Huttenhofer A., Wagner E.G.H. (2003) Rnomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucl. Acids Res*., **31**, 6435–6443.

Zaychikov E., Denissova L., Meier T., Götte M., Heumann H. (1997) Influence of Mg^{2+} and temperature on formation of the transcription bubble. *J. Biol. Chem*., **272**(4), 2259–2267.

ANALYSIS OF THE NUCLEOTIDE CONTEXT OF HIGHER PLANT MITOCHONDRIAL mRNA EDITING SITES

***Vishnevsky O.V.*[*1, 2], *Konstantinov Yu.M.*[3]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia; [3] Siberian Institute of Plant Physiology and Biochemistry, SB RAS, Irkutsk, 664033, Russia
[*] Corresponding author: e-mail: oleg@bionet.nsc.ru

Key words: mitochondrial mRNA editing sites, oligonucleotide motifs

SUMMARY

Motivation: C→U deamination ranks among the most widespread mechanisms of mitochondrial mRNA editing in higher plants. In the overwhelming majority of cases, editing affects the first and second positions of codons and results in "correction" of the codon, replacement in the amino acid sequence, and synthesis of normally functioning proteins. Experimental studies have shown that the 5'- and 3'-regions flanking an editing site are essential for precise and efficient editing. Nevertheless, no significant motifs have been found in the surrounding regions, and the editing mechanism still remains a mystery.

Results: We analyzed editing sites in mitochondrial mRNAs of *Arabidopsis thaliana* by the methods of region-specific degenerate oligonucleotide motifs and trinucleotide weight matrices. Significant oligonucleotide regularities were detected in the region [–50;50] with respect to editing sites. As shown by the jackknife method, these signals can be important for editing. However, it was shown that the features detected were insufficient for efficient recognition of editing sites in higher plant mitochondrial mRNAs.

Availability: http://wwwmgs2.bionet.nsc.ru./argo/.

INTRODUCTION

In spite of the fact that mitochondrial mRNA editing is widespread, its mechanisms in animals and plants are fundamentally different. In animal mitochondria, editing involves specific small RNAs referred to as guide RNAs, which have a clearly recognizable site of binding to the mRNA to be edited (Blum *et al.*, 1990). In contrast, there is no evidence for the involvement of guide RNAs in mRNA editing in higher plant mitochondria. No distinct signals directing editing in higher plants have been found either. It has been shown that the distribution of nucleotide frequencies in the immediate vicinity of an editing site is significantly nonrandom (Covello, Gray, 1990). There is experimental evidence (Takenaka *et al.*, 2004) that mutations in the [-40;-35] region reduce the efficiency of editing, whereas mutations in the [–15;-1] region entirely disrupt the process.

We analyzed editing sites by search for sets of degenerate region-specific oligonucleotide motifs. This method recognizes imperfect conserved signals within a reasonable time range without prior alignment. The analysis revealed significant oligonucleotide motifs in various stretches of the 5'- and 3'- regions flanking the edition sites.

METHODS AND ALGORITHMS

Sequences of 360 editing sites of Arabidopsis thaliana mitochondrial mRNAs were examined. Stretches of 101 bp in length were considered within the [–50;+50] region with respect to editing sites in mRNA coding regions.

To distinguish regularities specific for editing sites from frequency features of various reading frames, we divided the sample of editing sites into three subsamples with regard to the edition position in the reading frame. For each reading frame, the sample of unedited 101 bp-long stretches obtained from the same mRNAs and centered by a C was used as negative controls.

Search for degenerate oligonucleotide motifs was performed with the ARGO_Motifs program (Vishnevsky, Kolchanov, 2005). This algorithm involves clusterization of similar perfect oligonucleotides present in different sequences under study by an iteration method in the extended IUPAC code.

The resulting motif is considered significant if it meets the following requirements:

$$\left.\begin{array}{l} F > f_0 \\ P(n,N) < p_0 \\ Q < q_0 \end{array}\right\},$$

where F is the proportion of editing sites containing the motif; P(n,N), the binomial probability of the random occurrence of the motif in the window in ≥ n sequences of N; Q, the proportion of the sequences of the negative sample containing the motif; and f_0, p_0, and q_0 are threshold values.

The editing site recognition function R was estimated by the ARGO_Viewer method (Vishnevsky, Kolchanov, 2005), based on the comparison of motif frequency and distribution in a sequence under consideration and in sequences of edition sites in the training sample.

The positional context of editing sites was estimated by means of three-nucleotide weight matrices, which take into account local relationships between neighboring nucleotides. Positional weights were calculated as:

$$w_{b,k} = \log_2\left(\frac{f_{b,k}}{e_{b,k}}\right),$$

where $f_{b,k}$ is the frequency of the occurrence of trinucleotide b at position k in the sample of editing sites; $e_{b,k}$ is the frequency of its occurrence in the sample of nonsites.

The score of an unknown sequence of length L in the course of its analysis with the weight matrix was calculated as the sum of weights of corresponding positions.

$$W = \sum_{k=1..L} w_{b,k}.$$

We proposed an integrated approach to recognizing editing sites in higher plant mitochondrial mRNAs. A sequence S was considered to be such a site if the value of the recognition function T(S) = 1.

$$T(S) = \begin{cases} 1, R(S) > r_0 \_ and \_ W(S) > w_0 \\ 0, R(S) < r_0 \_ or \_ W(S) < w_0 \end{cases}$$

Here r_0, and w_0 are boundary values.

Samples constructed for three reading frames were used for training of recognition methods. Recognition quality was estimated in control sequences by the jackknife method. For this purpose, 30% of sequences were randomly chosen from the positive and negative samples to form control samples in the evaluation of recognition error. The remaining sequences were used for training. The procedure was performed in 100 replications to obtain mean error values. The value of the recognition function threshold yielding the minimum value of errors of type 1 (false negative) and of type 2 (false positive) was taken as optimal.

$$E = \frac{E_1 + E_2}{2},$$

where E_1 is the type 1 error and E_2 is the type 2 error.

IMPLEMENTATION AND RESULTS

Detection of degenerate oligonucleotide motifs

We applied the ARGO_Motifs method for context analysis of editing site surroundings. The search was performed in a 30 bp scanning window moving at a pace 15 bp at $p_0 = 10^{-14}$, $f_0 > 10$ %, $q_0 < 10$ %. The analysis performed in three reading frames showed that the edition sites of the first three frames contained significant motifs in both 5' and 3'-regions (Table 1). The absence of significant motifs from the third frame appears to be related to the critically small volume of the sample. Motifs obtained for the first and second frames are exemplified in the table. Note that the parameters and distributions of motifs found in both frames (Fig. 1) are similar. Approximately 80 motifs were found in the first and second frames. Their significance P varied within 10^{-14} to 10^{-21}, their frequency F in the site sample varied from 20 to 31 %, and in the sample of unedited positions Q, from 3 to 9 %. For example, motif **YTYYNTKT**, found in the [-35: -5] region with reference to the editing site in the first reading frame occurs in 31 % of editing-site sequences and only in 9 % of the mRNA sequences containing no editing sites.

Table 1. Parameters of the most significant oligonucleotide motifs found in the vicinity of editing site regions in the first and second reading frames

Motif	Location in the site sample	Occurrence in the site sample	Occurrence in the unedited mRNA sample	Probability of random presence of the motif
		First reading frame		
ATTYYNNT	-50: -20	0.28	0.08	10^{-18}
YTYYNTKT	-35: -5	0.31	0.09	10^{-20}
TTYCYNNT	-20: +10	0.31	0.09	10^{-20}
TYNYTCBK	-5: +25	0.35	0.09	10^{-20}
YTYNTTYT	+10: +40	0.25	0.08	10^{-19}
		Second reading frame		
TTTBTWWD	-50: -20	0.18	0.07	10^{-17}
HTWYKDTG	-35: -5	0.23	0.08	10^{-18}
WYTCVWNT	-20: +10	0.27	0.09	10^{-20}
YCNWWTCW	-5: +25	0.24	0.07	10^{-20}
TKNSAWWT	+10: +40	0.22	0.06	10^{-18}

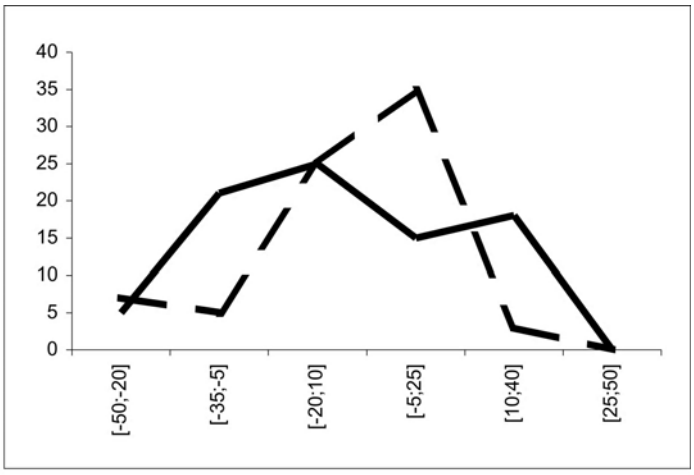


Figure 1. Distribution of the number of degenerate oligonucleotide motifs with reference to the editing site. X-axis, the number of motifs in the window; Y-axis, location of motifs. The solid line corresponds to the first reading frame and the dashed line, to the second frame.

Construction of the positional weight matrix

Analysis of contrasting weight matrices for the first and second reading frames showed that both of them contained trinucleotides whose frequency at certain positions of editing sites and unedited mRNAs differs significantly. Examples of such contrasting trinucleotides are shown in Table 2. They are either underrepresented by a factor of ≥ 8 ($w < -3$) or overrepresented by a factor of ≥ 4 at certain positions of editing sites in comparison with unedited mRNAs for both reading frames.

Table 2. Mean positional weights of trinucleotides underrepresented ($w < -3$) or overrepresented ($w > 2$) at a certain position of an editing site in both the first and second reading frames

Pos.	AUA	GAG	GGA	GGG	GGC	GCG	CAG	CGA	CGC	UCG	GUA
-3	-4.4	-4.2	-4.2	-3.2		-3.1		-3.6			
-2					-5.4				-5.3		
-1										2.2	
0											
1							-3.8				2.2

DISCUSSION

Our oligonucleotide analysis revealed many significant motifs in the 5'- and 3'-regions flanking editing sites. For both reading frames, the majority of significant motifs were found either in the immediate vicinity of the editing site or in the 5'-flanking region (Fig. 1). The motifs detected for two frames proved to be very similar. Most of them are polypyrimidine tracts dominated by U. It is worth noting that slightly fewer motifs were found for the second frame in the stretch [–35;–5] than in [–50;–20], which is in agreement with data reported in (Takenaka *et al.*, 2004), where it was shown that the [–35;–15] stretch was less significant for operation of an editing site than the neighboring stretches.

Analysis of the trinucleotide weight matrix shows that the local context of a mitochondrial mRNA editing site has notable constraints. Of importance is not only

presence but also absence of some short oligonucleotides. Our results confirm the experimental data (Choury *et al.*, 2004) that the presence of G at position –1 can entirely inhibit edition, whereas the elevated rate of UCG at position –1 is in agreement with the known high frequency of U there. In addition, the restrictions imposed on sites are entirely asymmetrical. Virtually all of them are related to the 5'-region with reference to the editing site.

It has shown that use of local regularities of editing sites is insufficient for their reliable recognition (Gray, Covello, 1993). To recognize editing sites in higher plant mitochondrial mRNAs, we proposed a combined approach, which involves both clearly located features, considered by the trinucleotide matrix method, and scattered motifs, detected by the ARGO package. We estimated the efficiency of this approach by the jackknife method. The least mean error values E for both reading frames considered equaled 0.27. The positional matrix best described the local context of a site, and oligonucleotide motifs, its distant features. Moreover, the method trained for recognizing editing sites at the first position of a frame recognized sites at the second position with the same mean error, 0.27, and vice versa. These data, taken together with the data on the similarity of motifs and their distribution, point to a similarity between the nucleotide contexts of editing sites in the first and second frames and, as a consequence, similar mechanisms of their operation.

Thus, we conclude that the regularities found in our study are essential for efficient edition of higher plant mitochondrial mRNAs. However, it is obvious that these context regularities are not quite sufficient for this process. It is known that the secondary structure of an mRNA is important for operation of its editing site and improves the quality of its recognition (Cummings, Myers, 2004). It should be noted that the least recognition errors were reported in (Mower, 2005), where the authors invoked for recognition data on the positional amino acid conservedness among species. Obviously, this method can be applied only to genes having many homologs in other plant species. It is reasonable to suggest that taking into account secondary structures regularities, comparative analysis in addition to degenerated oligonucleotide motifs will improve editing site recognition.

ACKNOWLEDGEMENTS

The work was supported by the RFBR (grants Nos 03-04-48829-a and 06-04-49556); SB RAS (integration projects Nos 5.1, 6, 47); Project “Evolution of molecular-genetic systems: computer analysis and modelling” of the RAS Presidium program “Biosphere origin and evolution” (#10104-34/P-18/155-270/1105-06-001/28/2006-1); Project “Computer modeling and experimental constructing of gene networks” of the RAS Presidium program of molecular and cell biology; State Contract with the Federal Agency for Science and Technology “Identification of potential targets for novel medicinal drugs based on reconstructed gene networks”, Frontiers in Genetics “Living Systems”; Innovation Project It-CP.5/001 “Development of software for computer modeling and design in postgenomic systems biology (*in silico* systems biology)” from the Federal Agency of Science and Innovation.

REFERENCES

Blum B., Bakalara N., Simpson L. (1990) A model for RNA editing in kinetoplastid mitochondria: “Guide” RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell*, **60**(2), 189–198.

Choury D., Farré J.C., Jordana X., Araya A. (2004) Different patterns in the recognition of editing sites in plant mitochondria. *Nucl. Acids Res*., **32**, 6397–6406.

Covello P.S., Gray M.W. (1990) Differences in editing at homologous sites in messenger RNAs from angiosperm mitochondria. *Nucl. Acids Res*., **18**, 5189–5196.

Cummings M.P., Myers D.S. (2004) Simple statistical models predict C to- U edited sites in plant mitochondrial RNA. *BMC Bioinformatics*, **5**, 132.

Gray M.W., Covello P.S. (1993) RNA editing in plant mitochondria and chloroplasts. *FASEB J*., **7**, 64–71.

Mower J.P. (2005) PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics*, **6**(1), 96.

Takenaka M., Neuwirt J., Brennicke A. (2004) Complex cis-elements determine an RNA editing site in pea mitochondria. *Nucl. Acids Res*., **32**, 4137–4144.

Vishnevsky O.V., Kolchanov N.A. (2005) ARGO: a web system for the detection of degenerate motifs and large-scale recognition of eukaryotic promoters. Nucl. Acids Res., **33**, Web Server issue W417–W422.

TRANSCRIPTION FACTOR BINDING SITES RECOGNITION BY THE REGULARITIES MATRICES BASED ON THE NATURAL CLASSIFICATION METHOD

***Vityaev E.E.*[*1, 2], *Lapardin K.A.*[2], *Khomicheva I.V.*[3], *Levitsky V.G.*[2, 3]**
[1] Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia;[2] Novosibirsk State University, Novosibirsk, 630090, Russia; [3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail : vityaev@bionet.nsc.ru

Key words: knowledge discovery and data mining, machine learning, eukaryotic promoter recognition, transcription factor binding sites

SUMMARY

Motivation: Numerous principles of constructing classifications are currently known. We propose the definition of "natural" classification and based on the definition a principally new approach to the classifications of nucleotide sequences.

Results: A method for constructing the "natural" classification, algorithm, and software system DNANatClass have been developed. As the application result we propose the regularities matrices describing SF1 and EGR1 transcription factor binding sites.

Availability: Scientific Discovery website: http://www.math.nsc.ru/AP/ScientificDiscovery.

INTRODUCTION

Position weight matrix is the most common method for the transcription factor binding sites (TFBSs) recognition. In this paper we present the regularities matrices that arise from the concept of natural classification in its application to the nucleotide sequences. The concept of natural classification was investigated and developed in the previous papers (Vityaev, 1983; Vityaev, Kostin, 1992; Vityaev *et al*., 2002). The main property of the regularities matrices is that each of the nucleotides A, T, G, C in each position of the matrix is characterized by its regularities connecting it with nucleotides in other positions, whereas the weight matrices estimates the contribution of each nucleotide taken separately without any interconnectivity.

Numerous principles of constructing classifications are currently known. The classifications are based on the hypothesis of compactness and various measures of closeness in a feature space, on resemblance of standards, supertargets, various criteria of classification quality and quality functionals, separation of distribution mixtures, etc. (Classification and Clustering, 1977). In contrast to the above-listed classifications the objective of the "natural" classification is discovering the laws of nature. There are different definitions of the natural classification that were done by the naturalists in different times (see overview in Zabrodin, 1981). We propose the definition of the "natural" classification that is in accordance with the definitions of naturalists: "Objects should be divided into classes in accordance with the regularities satisfied by the objects. Objects of one class should obey one group of regularities, and objects of different classes should obey different groups of regularities. Objects of one

class should also possess some integrity which is understood as mutual prediction of object properties" (Vityaev, 1983).

METHODS AND ALGORITHMS

The following method realizes the above definition of the natural classification and includes three steps: regularities determining, classes formation and recognition (Vityaev *et al*., 2006a).

1. Regularities discovery.

Definition 1. The rule $(P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk} \Rightarrow P^{\varepsilon 0}{}_{i0j0})$ is the *probabilistic law* if and only if:

1) $\mu(P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk}) > 0$;

2) $\mu(P^{\varepsilon 0}{}_{i0j0}/P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk}) > \mu(P^{\varepsilon 0}{}_{i0j0}/P^{\varepsilon 1}{}_{i1j1}\&\ldots\hat{}\ldots\hat{}...\&P^{\varepsilon k}{}_{ikjk})$, where ...^…^..

means the absence of one or more predicates in the premise of the rule, and conditional probability is defined as follows: $\mu(P^{\varepsilon 0}{}_{i0j0}/P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk}) = \mu(P^{\varepsilon 0}{}_{i0j0}\&P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk})/ \mu(P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk})$.

For the predicate $P^{\varepsilon 1}{}_{i1j1}$ the index i1 means the position number, j1 means one of the nucleotide {A,T,G,C}, ε = 0/1 means that the predicate has/hasn't the negation. For example, the predicate $P^{1}{}_{ikA}$ means that in the position ik there is the nucleotide A. Let $\mu(\varphi) = \mu(P^{\varepsilon 0}{}_{i0j0}/P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk})$ be the conditional probability of the rule φ. Given the sample of the sequences we discover the set of regularities F. By the estimation of regularity we mean the value $\mu^{\beta}(\varphi) = -\ln(1-\mu(\varphi))$ calculated with confidence level β.

2. Classes discovery. Let us define the criterion of regularities interconnection. By the tuple of properties values $x_{s1},...,x_{sm}$ we call the set $\{Y_{s1},...,Y_{sm}\}$, $Y_{st} \subset I_{st}$, $Y_{st} \neq \varnothing$, t = 1,...,m, I_{st} – the set of all values of the feature st. We designate that the regularity $(P^{\varepsilon 1}{}_{i1j1}\&...\&P^{\varepsilon k}{}_{ikjk} \Rightarrow P^{\varepsilon 0}{}_{i0j0})$ *is applied* to the set $\{Y_{s1},...,Y_{sm}\}$, if $\{i_0,i_1,...,i_k\} \subset \{s1,...,sm\}$ and also $x_{itjt} \in Y_{it}$ if εt = 1 and $(x_{itjt} \notin Y_{it})$ if ε = 0, t = 1,...,k. If the regularity is applied to the set $\{Y_{s1},...,Y_{sm}\}$ and the conclusion of the rule $P^{\varepsilon 0}{}_{i0j0}$ is fulfilled for that set ($x_{i0j0} \in Y_{i0}$ if ε = 1 and $x_{i0j0} \notin Y_{i0}$ if ε = 0), then we say that the regularity is *satisfied* for that set, but if conclusion is not fulfilled, then we say that the regularity is *falsified* for that set. By the criterion of regularities interconnection on the set $\{Y_{s1},...,Y_{sm}\}$ we designate the value:

where Π is the set of satisfied regularities, and O is the set falsified regularities.

Definition 2. By the *class* we call the set $\{Y_{s1},...,Y_{sm}\}$, for which the criterion Γ

$$\Gamma(\{Y_1,...,Y_m\}) = \sum_{\varphi \in \Pi} \mu^{\beta}(\varphi) - \sum_{\varphi \in O} \mu^{\beta}(\varphi)$$

reaches the local maximum.

The set $\{Y_{s1},...,Y_{sm}\}$ cay be presented as the matrix. For example the sequence [A][A][C][A][G][C][T][A][C][A][G][G][T][A][A][G][G][G][C][T] cay be presented as matrix $M(Y_{s1},...,Y_{sm})$:

A	1	1	0	1	0	0	0	1	0	1	0	0	0	1	1	0	0	0	0	0	1
T	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
G	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	1	1	1	1	0	0
C	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0

In addition to the matrix $M(Y_{s1},...,Y_{sm})$ we define the regularity matrix $R(Y_{s1},...,Y_{sm})$ as the matrix of predictions of the cells of the matrix $M(Y_{s1},...,Y_{sm})$ by regularities. The sum of values of regularity matrix $R(Y_{s1},...,Y_{sm})$ is equal to the criterion $\Gamma(\{Y_{s1},...,Y_{sm}\})$. Also we use the involvement matrix $I(Y_{s1},...,Y_{sm})$ to show the involvement of all predicates of the regularities in there interconnection, which have estimation $\mu^{\beta}(\varphi)$ for each predicate of the regularity.

3. Recognition. Given the control set B of sequences, class $O_i = \{Y_{i1}, ..., Y_{in}\}$, and the set of regularities F we can recognize the positive and control samples by calculating the score $\Gamma(\{Y_{i1}, ..., Y_{in}\})$ for every training and control sequence. When we define some threshold of the score, we can calculate the true/false positive rates for the training and control sets.

IMPLEMENTATION AND RESULTS

For the TFBSs recognition we have chosen the samples of sites SF1, EGR1. The train data sets were extracted from the TRRD database (Kolchanov *et al.*, 2002). We added to the positive samples the sets of randomly generated sequences, which were generated with the same frequencies as for the positive samples. The number of randomly generated sequences was ten times more then the number of positive sequences. Then using that mixed sample we performed the classification of the whole data and discovered the class(es) for the positive samples. Exactly the one class was discovered for the SF1, EGR1 sites samples.

The negative control sample was randomly generated with the frequencies as in the positive samples. For the recognition of the positive and control sequences we first performed the classification of that samples. Then the score Γ was calculated for the positive and control sequences that were classified as belonging to the class. We defined the threshold for which the 50 % of positive sequences were recognized as belonging to the class. With this threshold we calculated the false positive rate. The more detailed description of results is depicted in the following table.

Table 1. Table of results

Site	# positive sequences	#negative control sequences	# of regularities belonging to	# of classified positive	# of classified negative	the score	# of recognized positive sequences	# of recognized negative	false positive rate
SF1	54	100000	1670	54	81940	3900	25	2	2/100 000
EGR1	22	110000	789	16	25502	900	7	0	0/110 000

The class [T/C][C][A][A][G][G][T/C][C][A][G] was discovered for the SF1 site, where [T/C] means that on the first place there can be one of two nucleotides T or C. The class [G][C][G][G][G][G][G][CA][G][G] was discovered for the **EGR1** site.

The regularity matrix R([TC][C][A][A][G][G][TC][C][A][G])

A	0.00	0.00	0.00	0.00	17.92	1512.67	0.00	0.00	0.00	0.00
T	0.00	484.95	643.73	481.73	421.14	872.68	0.00	0.00	0.00	0.00
G	154.92	0.00	2.31	61.84	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	4.13	9.89	103.06	634.36	0.00	0.00	0.00	0.00

The regularity matrix R([G][C][G][G][G][G][G][CA][G][G]).

A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	447.36	0.00	-47.98	-15.30	-2.69	0.00	515.56	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

DISCUSSION

Further we plan to improve the method and use it in cooperation with the ExpertDiscovery method (Vityaev *et al*., 2006b). We can discover the complex signals by the ExpertDiscovery system and use them as ordinary properties in the classification system DNANatClass.

ACKNOWLEDGEMENTS

The work is partially supported by the Russian Foundation for Basic Research No. 05-07-90185-v, Scientific Schools grant at the President of the Russian Federation No. 4413.2006.1, Innovation project IT-CP.5/001 “Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)”.

REFERENCES

Classification and Clustering (1977) Van Ryzin J. (ed.), Academic Press, New York.

Kolchanov N.A. *et al*. (2002) Transcription Regulatory Regions Database, (TRRD): its status in 2002. *Nucl. Acid Res*., **30**, 312–317.

Vityaev E.E. (1983) Classification as a determination of groups of objects that satisfy different sets of consistent regularities. *Comp. Syst*., **99**, 44–50 (in Russ.).

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., part 4) 107–115 (in Russ.).

Vityaev E.E. *et al*. (2002) Natural classification of nucleotide sequences. *Proc. of the Third International Conference On Bioinformatics of Genome Regulation and Structure,* BGRS’2002, Novosibirsk, Russia, July 14-20, ICG, Novosibirsk, **3**, 197–199.

Vityaev E.E. *et al*. (2006a) Natural classification and systematic as the laws of nature. Analysis of structural regularities (*Comp. syst*. No. 174), Novosibirsk (in Russ.) (in press).

Vityaev E.E. *et al*. (2006b) Software for analysis of gene regulatory sequences by knowledge discovery methods. In Kolchanov N., Hofestaedt R. (eds), *Bioinformatics of Genome Regulation and Structure II. Springer Science+Business Media*, Inc., pp. 491–498.

Zabrodin V.Yu. (1981) Criteria of naturalness of classifications. NTI, ser. 2.

ROLES OF CODON BIASES AND POTENTIAL SECONDARY STRUCTURES IN mRNA TRANSLATION OF UNICELLULAR ORGANISMS

Vladimirov N.V.*, *, Likhoshvai V.A., Matushkin Yu.G.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: nikita@bionet.nsc.ru

Key words: translation, ORF, codon usage, secondary structure

SUMMARY

Motivation: Correlation of gene expression with the degree of codon bias is known in many unicellular organisms. However, in a number of organisms such correlation is absent. Recently we have shown that consideration of inverted complementary repeats within open reading frames (ORFs) is necessary for proper estimation of translation efficiency (Likhoshvai, Matushkin, 2002).

Results: An algorithm for estimation of potential ORF expression in an organism using its genome sequence is proposed. The potential ORF expression is estimated using the elongation efficiency index (EEI). Computation is based on estimation of ORF elongation efficiency considering three key factors: codon bias, average number of inverted repeats within coding sequence, and free energy of potential stem-loop structures. Quantitative translational characteristics of 240 unicellular organisms (213 bacteria, 22 archaea, and 5 eukaryota) have been computed. Five potential evolutionary strategies of translational optimization are defined among studied organisms. A significant difference of preferred translational strategies between Bacteria and Archaea has been revealed.

Availability: http://wwwmgs2.bionet.nsc.ru/mgs/programs/eei-calculator/.

INTRODUCTION

Elongation is the most energy- and time-consuming stage of mRNA translation. Therefore, high level of gene expression requires high rate of elongation. In many unicellular organisms this is achieved by non-uniform usage of synonymous codons, with preferences for a subset of "optimal" codons in highly-expressed genes (Sharp, Li, 1987; Andersson, Kurland, 1990). The subset of preferred codons in such organisms has high relative concentrations of cognate tRNAs (Gouy, Gautier, 1982; Ikemura, 1985). Codon bias increases translation rate because preferred codons tend to be translated more rapidly than synonymous alternatives (Kurland, 1991). Translational codon bias is correlated with gene expression level in many prokaryotes and some eukaryotes (Gouy, Gautier, 1982; Ikemura, 1985; Duret, Mouchiroud, 1999).

Though codon bias indices like CAI (Sharp, Li, 1987) are good predictors of gene expression in *E. coli*, *B. subtilis* and many other organisms, they are not useful for organisms which do not have strong translational codon bias. This set of species includes *Helicobacter pylori* (Lafay *et al*., 2000), *Borrelia burgdorferi* (Lafay *et al*., 1999) and others. In these species codon bias is not correlated with gene expression, and indices like CAI can not be applied to gene expression prediction.

This problem prompted the authors to consider the negative influence of RNA secondary structures on the rate of elongation (Likhoshvai, Matushkin, 2002). To estimate this influence, authors proposed calculation of the average number of local inverted complementary repeats within ORF, which may form local stem-loop structures.

METHODS AND ALGORITHMS

Genome sequences. The flat files of 240 complete genome sequences of unicellular organisms were retrieved from NCBI GenBank FTP.

Local complementarity index. In order to estimate the influence of local secondary structures on elongation efficiency, authors proposed local complementarity index (LCI) in two forms (LCI_L and LCI_E), with and without taking into account the free energy of potential stem-loops (Likhoshvai, Matushkin, 2002). The free energy of stem-loop structures was calculated according to the nearest-neighbor model (Turner, Sugimoto, 1988).

Elongation Efficiency Index. After calculation of $LCI_L(j)$ and $LCI_E(j)$ indices for each ORF_j, an iterative algorithm ranks ORFs according to their EEI values. The *EEI(j)* value reflects the relative average elongation rate of one codon in ORF_j:

$EEI(j)=K/(w_1T_a(j)+w_2T_e(j))$, where K is a scale constant, w_1 = (0 or 1), and w_2 = (0 or 1) are weight coefficients.

The first term T_a evaluates the codon bias. The second term, $T_e(j)$, estimates the mean time required for translocation (Likhoshvai, Matushkin, 2002).

Weight coefficients w_1, w_2 have values 0 or 1, and LCI(j) may be of two types, LCI_L(j) or LCI_E(j), so there are five forms of EEI:

1) EEI_1 (A), $w_1 = 1$, $w_2 = 0$, no *LCI*. Only codon bias is considered, and secondary structures are neglected.

2) EEI_2 (LCI_L), $w_1 = 0$, $w_2 = 1$, $LCI(j)=LCI_L(j)$. Codon bias is neglected. Only the number and lengths of secondary structures are considered.

3) EEI_3 (LCI_E), $w_1 = 0$, $w_2 = 1$, $LCI(j)=LCI_E(j)$. Codon bias is neglected. Only the number and free energies of secondary structures are considered.

4) EEI_4 (A-LCI_L), $w_1 = 1$, $w_2 = 1$, $LCI(j)=LCI_L(j)$. Both codon bias and number of secondary structures with account of their lengths are considered.

5) EEI_5 (A-LCI_E), $w_1 = 1$, $w_2 = 1$, $LCI(j)=LCI_E(j)$. Both codon bias and number of secondary structures with account of their energies are considered.

Estimation of correlation between EEI and gene expression. To estimate correlation of EEI with gene expression levels, we used ribosomal genes as a set of highly expressed genes in unicellular organisms. They were used as markers to evaluate the ability of EEI to predict gene expression level.

For each of the five EEI types we calculated a pair of ($M\pm R$) – the *relative elongation efficiency* of ribosomal genes. M is the normalized mean of positions of ribosomal genes among all genes ranked by EEI, and R is the normalized standard deviation of positions. The normalization of (M,R) consists in linear transformation, so that M becomes symmetrically scaled relatively to zero: $-100 < M < 100$, and $0 < R < 100$.

All studied organisms may be classified into five translational groups according to the leading type of translational index M_i. An organism falls into one of five groups according to the maximal value of M_i (and minimal R_i).

The statistical significance of $M_i\pm R_i$ realization was estimated using Monte Carlo simulations. The null hypothesis consists in uniform random distribution of ribosomal genes among other genes ordered by EEI. For most genomes the statistical significance is of order $p \cong 10^{-8}$–10^{-12}.

IMPLEMENTATION AND RESULTS

Using the EEI values of genes encoding ribosomal proteins, we show that EEI is highly correlated with gene expression in 240 unicellular organisms.

Bacteria. We computed translational characteristics for 213 bacterial genomes available at NCBI GenBank database on August 1, 2005. The values of the (M_i,R_i) pairs for some bacteria are shown in Table 1. The cumulative diagram of bacterial genome distribution over five translational groups is shown in Fig. 1. Most bacterial genomes fall into Groups 1 and 4.

Table 1. Relative elongation efficiency of ribosomal genes in some bacterial genomes

Genome	Num of ribos. genes	Num of all genes	$M_1 \pm R_1$	$M_2 \pm R_2$	$M_3 \pm R_3$	$M_4 \pm R_4$	$M_5 \pm R_5$	Group
Escherichia coli K12	59	4270	89±33	23±63	23±59	75±42	61±56	1
Mycoplasma hyopneumoniae	48	691	-56±55	73±34	64±44	41±53	47±51	2
Nitrosomonas europaea	55	2573	-52±60	18±67	63±47	-6±69	57±49	3
Borrelia burgdorferi	53	848	-14±60	57±47	57±52	65±41	53±50	4
Pseudomonas putida KT2440	54	5350	74±26	60±33	67±27	86±17	92±13	5

$-100<M_i<100$, $0<R_i<100$ (i = 1, .., 5) are normalized mean and standard deviations of ribosomal genes positions among other genes ranked by EEI_i values. The highest M_i with corresponding R_i (*i* = 1, .., 5) are shadowed.

Archaea. We computed translational characteristics for 22 archaeal genomes.

The cumulative diagrams of distributions for bacterial and archaeal genomes over five translational groups are shown in Fig. 1. The most numerous group is 4 (15 organisms, 68 % of the total number). Less numerous is Group 2 (4 organisms).

Unicellular Eukaryota. We have computed translational characteristics for 5 genomes of unicellular eukaryotes: *S. cerevisiae*, *S. pombe*, *E. cuniculi*, *G. theta*, and *P. falciparum*. Both yeast species and *E. cuniculi* fall into Group 1; *G. theta*, into 5; and *P. falciparum* into Group 2.

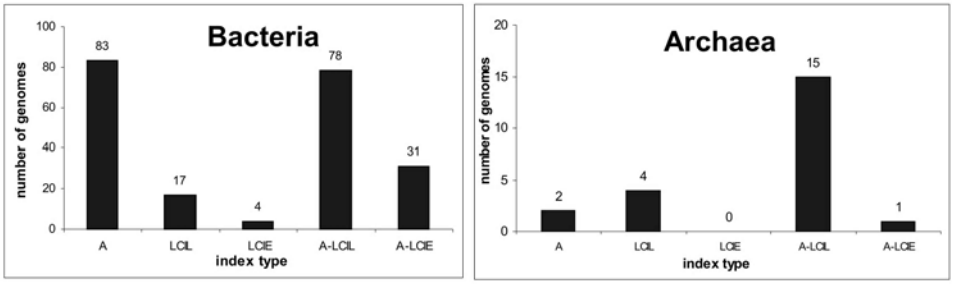


Figure 1. Distribution of 213 bacterial and 22 archaeal genomes over 5 translational groups.

DISCUSSION

The results of computations imply high correlation between EEI and gene expression in 240 unicellular organisms. This indicates that elongation efficiency in most known unicellular organisms may be determined by two key factors – frequencies of preferred

codons and avoidance of nonspecific secondary structures. The EEI values of ribosomal genes are also high in organisms like *H. pylori*, where traditional codon indices (CAI) do not correlate with gene expression. We assume that EEI may be used for prediction of gene expression in such organisms, and other unicellular organisms as well.

Domains of Bacteria and Archaea substantially differ in distributions by translational groups. In the translational characteristics of 68 % of the considered Archaea species, both codon bias and potential hairpins with account of their length are the key factors (Group 4). In contrast, the most populated among bacterial species is Group 1 (39 %), where translational efficiency is determined only by codon biases, and secondary structures are neglected. Such high importance of secondary structures in translation of Archaea may be associated with extreme temperature and pH conditions of their environment.

The proposed approach allows to estimate whether gene expression in an organism is correlated with selection towards preferred codons (high M_1) and/or selection against secondary structures (high M_2, .., M_5). This knowledge may be useful in planning transgenic studies with recently sequenced unicellular organisms.

ACKNOWLEDGEMENTS

The work was supported by the Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49068, 05-07-90274, 06-04-49556), Russian Ministry of Industry, Science and Technology (No. 43.073.1.1.1501), Siberian Branch of the Russian Academy of Science (project No. 10.4), Integration Project No. 119, and State contract #10104-34/P-18/155-270/1105-06-001/28/2006-1 and NSF:FIBR (Grant No. EF–0330786).

REFERENCES

Andersson S.G.E., Kurland C.G. (1990) Codon preferences in free-living microorganisms. *Microbiol. Rev.*, **54**, 198–210.

Duret L., Mouchiroud D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. *Proc. Natl. Acad. Sci. USA*, **96**(8), 4482–4487.

Gouy M., Gautier C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucl. Acids Res.*, **10**, 7055–7070.

Ikemura T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.

Kurland C.G. (1991) Codon bias and gene expression. *FEBS Lett.*, **285**(2), 165–169.

Lafay B. *et al.* (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucl. Acids Res.*, **27**, 1642–1649.

Lafay B. *et al.* (2000) Absence of translationally selected synonymous codon usage bias in Helicobacter pylori. *Microbiology*, **146**, 851–860.

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

Sharp P.M., Li W.H. (1987) The codon adaptation index – a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acid Res.*, **15**, 1281–1295.

Turner D.H., Sugimoto N. (1988) RNA structure prediction. *Annu Rev. Biophys Biophys Chem.*, **17**, 167–192.

MODELING OF DATA BASE OF CONTEXT-DEPENDENT CONFORMATIONAL PARAMETERS OF DNA DUPLEXES

Vorobjev Y.N.**, *Emelianov D.Y.
Institute of Chemical Biology and Fundamental Medicine, SB RAS, Novosibirsk, 630090, Russia;
Novosibirsk State University, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: ynvorob@niboch.nsc.ru

Key words: molecular dynamics, DNA conformational dynamics, context dependent DNA conformational parameters

SUMMARY

Motivation: A context dependent conformational preference and deformability of DNA plays as a significant factor for a DNA site recognition by a DNA-binding regulatory and nucleosome proteins. An extraction of context-dependent conformational and deformability parameters from static X-ray structures of crystals of DNA duplexes includes artifacts due to a crystal packing. A molecular dynamic simulation of a large series of 14 base pairs DNA duplexes of different sequences in water solution at physiological conditions have been done. The context-dependent conformational parameters are extracted from the simulated trajectories of thermal fluctuations of three-dimensional structures of DNA duplexes . It is found that helical parameters of pair steps **TA, TG, CG** are anomalous and sharply neighbor-context dependent.

Results: molecular modeling provide data to expand bioinformatics data bases beyond the capacity of experimental methods and provide a new knowledge.

INTRODUCTION

Formation of the protein-DNA complex depends on two major factors: 1) structural affinity of the average three dimensional structure of DNA binding site, and 2) binding site deformability to make induced fit. A context dependent average conformational parameters of DNA and values of its thermal fluctuations can be itself an important data to make a prediction of a functional characteristics of a sequences. Therefore a reliable set of a context dependent conformational/deformability parameters of DNA can serve as a natural descriptors to describe a different biological properties of DNA sequences and its response on mutations. A complete set of context dependent conformational/deformability parameters is still unknown because analysis of experimental static crystal structures of DNA duplexes or protein/DNA complexes include artifacts due to crystal packing effects (Vorobjev, 2003). Molecular modeling of internal conformational dynamics of DNA duplexes due to thermal fluctuation is able to provide data of the context dependent average conformational parameters and its thermal fluctuations which defines a deformability. The molecular dynamics simulations of a large set of 14-base pairs DNA duplexes are performed in an aqueous solvent with neutralized counterions at physiological conditions. A dynamic average and value of thermal fluctuations of helical parameters for all pair nucleotide steps XY are extracted via statistical analysis of molecular dynamics data.

METHOD

1.1 Description of context dependent properties of DNA duplexes

A double stranded DNA duplex of A,G,C,T nucleotides can be represented as a linear sequence of ten types of pair-nucleotide steps **XY** (Vorobjev, 2003). The conformational parameters of pair nucleotide step of DNA duplex are shown in Fig. 1.

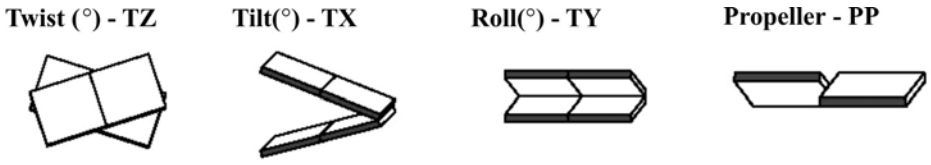


Figure 1. Helical parameters for pair base step of the DNA-duplex. Twist (°) (TZ) – rotation around Z-axis normal to the average base pair planes; Tilt(°) (TX) – opening angle between long axes of base, pairs-rotation around short axes of base pare; Roll(°) (TY) – opening angle between short axes of base pairs in the direction of minor (major) groove, rotation around long axes of base pare; Propeller (PP) – angle between base planes of the Watson-Crick base pair.

Helical parameters of the step **XY** are a function of the dinucleotide **XY** type and flanking context of the both sides, $\ldots X_{-3}X_{-2}X_{-1}\mathbf{XY}Y_1Y_2Y_3\ldots$. In general, the influence of the remote context on the step **XY** will decrease. Therefore the helical parameters dinucleotide step can be considered at different level of concretization of flanking sequences, i.e. simple **pair step** <**XY**> averaged over all flanking sequences, **quartet step** $\langle X_{-1}\mathbf{XY}Y_1\rangle$, etc. The most important helical parameter affecting overall global spatial structure of DNA is the **tilt** which control bending of DNA rod into major or minor groves.

1.2 Molecular dynamic simulations

Molecular dynamics simulations have been done with amber6 (URL:2001) program using param98 force-field parameter set. Simulation protocol consist of the next stages: 1) calculation of initial coordinates of duplex atoms in standard B-form; 2) short energy minimization in vacuum with distant dependent dielectric constant; 3) solvatation of the dna-duplex in the rectangular box with 9 Å distance from the nearest dna atom to the box side; 4) neutralization of the dna-duplex by Na^+ ions; 5) energy minimization of water and ion positions until grad(E) < 0.1 kcal/mol/ Å; 6) slow heating from 1 to 300 K in 10 ps with the soft harmonic restraint potential (the harmonic potential constant, K_h = 0.1 kcal/mol/ $Å^2$) for dna-duplex atoms; 7) final equilibration during 50 ps with the soft harmonic restraint potential for atoms of dna-duplex flanking base pairs; 8) productive molecular dynamics run of 1500 ps at T = 300K, P = 1 bar with PME for the long-range electrostatic forces and weak harmonic restraint potential (the harmonic potential constant, K_h = 0.02 kcal/mol/ $Å^2$) for dna-duplex flanking base pair and trajectory collection with 1 ps interval.

A typical fluctuation behavior of the base step helical parameters along a molecular dynamic trajectory of the DNA duplex are shown in Fig. 2. It can be seen that high frequency fluctuations (ps scale) have a large amplitude and describe a fast local conformational fluctuations. The average value, over the 50 ps window, shows quite smooth behavior with a period of slow fluctuation of about 500–600 ps. Therefore it can be concluded that trajectory of 1500 ps of length provides a reasonable amount of data to obtain an average and statistical fluctuation values for the helical parameters of DNA duplex. The molecular dynamic simulation of 150 DNA 14-base pairs duplexes have been performed according to described protocol. A series of instant structures (1 500 000 structures include 1800 pair steps) have been collected. The average value of helical parameters <P> and value of thermal fluctuations $\langle\Delta P^2\rangle^{1/2}$ at 300K have been calculated for the collected data base. The thermal fluctuation $\langle\Delta P^2\rangle^{1/2}$ defines a conformational rigidity (or deformability) $K_P = 2kT/\langle\Delta P^2\rangle$ of potential energy profile along conformational parameter P.

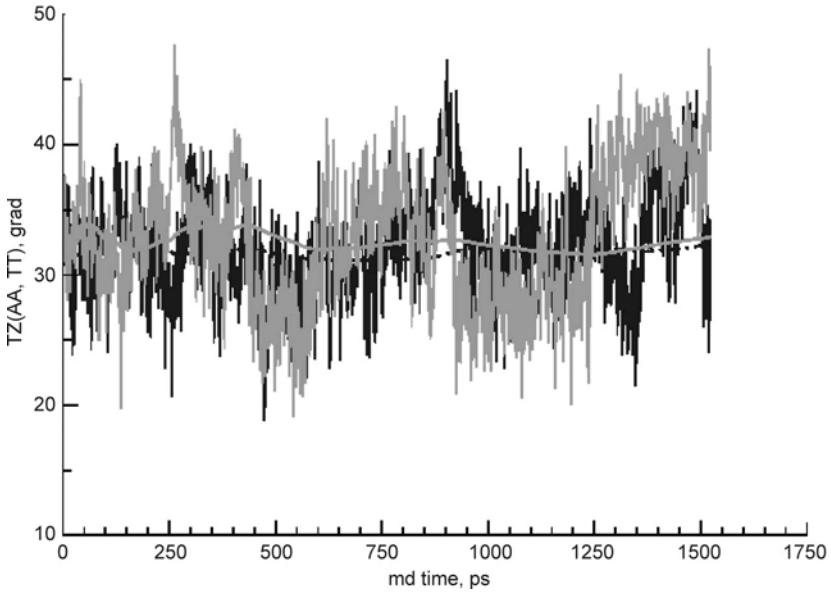


Figure 2. Fluctuation of the helical angle Twist (TZ) for the T4T5 (dark) and A7A8 (grey) nucleotide step of the dna14-1 duplex. Thick smooth lines are the average over 50 ps window.

RESULTS

It have been found that statistical accuracy of about 0.25 ° for the average value of helical parameters of dinucleotide step XY can be achieved for averaging over ~120 flanking sequences, Fig. 3.

Table 1 shows results of simulation of helical parameters for ten types of dinucleotide steps. It can be seen that the helical parameters for the dinucleotide XY steps depend on XY context.

DISCUSSION

Table 1 shows that three pair steps **TA, TG, CG** have large average **tilt angle <TY>** ~ -11°, compare to that of - 3° for average pair step **XY**. Therefore a special distribution of the TA,TG,CG in the DNA sequence can lead to a formations of the DNA rods macroscopically bended in major groove direction. The extent of fluctuations over flanking sequences consist of two groups, a) **neighbor context sensitive** four pair steps, namely **TA, TG, CG**, **AC** with large values of fluctuations of **tilt** and/or **twist** helical parameters ~ 5.5°, and group b) of **neighbor context insensitive** of six remaining pair steps having a value of fluctuations about 3.3°. The neighbor context sensitive pair steps should be analysed on the quartet level, i.e. $<X_1\mathbf{XY}Y_1>$.

The values of thermal fluctuations (and coupled deformability) of helical parameters at 300K are context insensitive. The propeller parameter is a characteristics of one complementary base pair which controls a relative positions of the DNA atoms donor/acceptors of H-bonds on the surface of minor/major groves. The **propeller** parameter has extreme values for **AT** complementary pair steps.

It can be noticed that the average values of helical parameters in the table 1 and values extracted from the X-ray structures of DNA duplexes (Bhattacharyya *et al.*, 1999; Vorobjev, 2003) have a low correlation coefficient. As shown in Ref. (Bhattacharyya *et al.*, 1999), a crystal field can considerably affect the average value of helical parameters of

DNA duplexes. The value of fluctuations of the helical parameters calculated on the data base of the X-ray crystal structures is probably insensitive on crystal field effects. As found in Ref. (Vorobjev, 2003) on the set of DNA duplex crystal consisted of 644 dinucleotide pair step, the larges fluctuations has the tilt parameter of pair steps TA, TG, CG, GC, AC in a good agreement with results of the presented modeling.

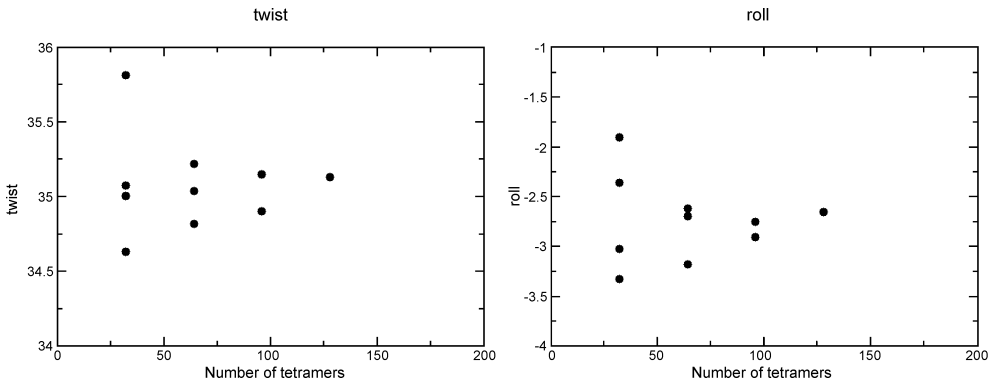


Figure 3. Dependence of average value of helical parameters **twist** and **roll** on the number of flanking sequences for the AA step.

Table 1. Equilibrium helical parameters and their thermal fluctuations over MD trajectory for ten dinucleotide steps and its average and standard deviations over flanking sequences

n pair step	Equilibrium values				Thermal fluctuations			
X-Y	**TX**	**TY**	**TZ**	**PP**	***TX***	***TY***	***TZ***	***PP***
5′-3′/5′-3′	$\langle\langle H\rangle_t\rangle_f$				$\langle SD[H]_t\rangle_f$			
	$SD[\langle H\rangle_t]_f$				$SD[SD[H]_t]_f$			
1 A-A/T-T	-3.6	-2.5	35.3	**-21.8**	6.4	10.2	4.7	15.4
	1.1	3.0	2.1	4.9	0.2	1.0	0.8	2.3
2 A-T/A-T	-0.2	2.4	33.2	**-19.8**	7.1	9.3	3.9	15.3
	0.9	3.2	2.2	4.8	0.2	1.2	0.9	2.1
3 A-G/C-T	-4.8	-2.7	34.6	-15.7	6.5	10.7	4.9	16.8
	1.2	3.6	2.5	5.5	0.3	1.0	1.0	1.9
4 A-C/G-T	-2.3	1.0	33.2	-14.6	6.9	9.8	4.6	16.6
	1.3	**5.5**	3.6	6.0	0.2	1.3	0.9	2.1
5 T-A/T-A	-0.4	**-10.2**	33.7	-15.2	6.1	11.6	6.1	17.7
	1.3	**4.9**	3.8	6.1	0.2	1.2	1.1	2.2
6 T-G/C-A	-1.0	**-11.7**	33.2	-11.0	5.9	11.5	6.4	17.8
	1.4	**5.6**	**4.8**	6.2	0.2	0.9	1.1	1.9
7 T-C/G-A	-1.0	-0.6	36.8	-13.9	6.4	10.6	5.0	17.0
	1.7	3.3	2.1	8.4	0.2	1.0	0.8	2.3
8 G-G/C-C	-0.5	-2.8	34.7	-8.7	6.2	10.7	5.4	17.7
	1.6	3.4	2.4	6.5	0.2	1.4	1.0	1.5
9 G-C/G-C	0.1	3.7	36.0	-10.3	7.1	10.2	4.9	17.9
	2.0	2.9	2.0	5.7	0.2	1.2	1.0	1.7
10C-G/C-G	-0.3	**-11.3**	34.8	-9.5	6.1	11.3	6.1	17.9
	1.9	**5.9**	**6.2**	7.1	0.4	1.1	1.5	1.6
all average	-1.2	-3.2	34.6	-13.4	6.5	10.6	5.2	17.0
	1.5	4.1	3.2	6.8	0.2	1.2	1.1	2.1

$\langle\langle H\rangle_t\rangle_f$ – average Helical parameter over **time snapshots** for particular XY pair step of DNA duplex and then it are averaged over all **flanking** sequences; $SD[\langle H\rangle_t]_f$ – standard deviation of time average helical parameter over all flanking sequences; $\langle SD[H]_t\rangle_f$ – standard deviation over time snapshots are averaged over all flanking sequences, it is equal to the average value of thermal fluctuations; $SD[SD[H]_t]_f$ – standard deviation over all flanking sequences for standard deviations over time snapshots.

ACKNOWLEDGEMENTS

The work was supported by the RFBR grant #05-04-48322 and interdisciplinary project #119 of the SB RAS.

REFERENCES

Bhattacharyya D., Kundu S., Thakur A.R., Majumdar R. (1999) Sequence directed flexibility of DNA and role of cross-strand hydrogen bonds. *J. Biomol. Struct. Dynamics*, **17**, 2, 289–299.

URL: 2001. AMBER6 Home page: http://www.amber.ucsf.edu/amber/index.html.

Vorobjev Y.N. (2003) In silico modeling and conformational mobility of DNA duplexes. *Mol. Biology. (Transl. from Rus.)*, **37**, 2, 210–222.

SELECTION OF INFORMATIVE SUBSET OF GENE EXPRESSION PROFILES IN PROGNOSTIC ANALYSIS OF TYPE II DIABETES

Zagoruiko N.G.*[*1], Kutnenko O.A.[1], Borisova I.A.[1], Kiselev A.N.[1], Ptitsyn A.A.[2]
[1] Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Pennington Biomedical Research Center, 6400 Perkins Rd, Baton Rouge LA 70808, USA, e-mail: PtitsinAA@pbrc.edu
[*] Corresponding author: e-mail: zag@math.nsc.ru

Key words: microarray phenotyping, feature selection, diagnosis, obesity, diabetes

SUMMARY

Motivation: High-density microarray data can be a rich source of information and play a key role of understanding the etiology of complex multifactorial disease such as Diabetes mellitus type II (DM2). However, selection of features important for diagnostic and prognostic purpose out of over 22000 transcripts represented on a microarray is a serious challenge. Additional complication comes from the fact that features represented in the microarray data are not independent, but intertwined in a complex network of relations, which itself is a subject of research.

Results: We have developed the algorithms of directed search GRAD and LAD. These algorithms allow selection of an informative subset of genes with account of their inter-dependence. The effectiveness of these algorithms has been tested in experiments with selection of informative genes related to the diagnostics of DM2. We introduce a new method of sorting the observed molecular phenotypes of patients' skeletal muscle on a scale ranging from most to least metabolically fit. The rank of a particular phenotype is highly correlated to the individual risk of developing DM2 and can be used for diagnostic and prognostic purpose.

Availability: http://compbio.pbrc.edu/pti.

INTRODUCTION

Data obtained in a microarray experiment represents a coarse-grained snapshot of expression for thousands of genes. This "molecular fingerprint" is bound to reflect not only routine housekeeping activity, but also patterns specific for particular organs, tissues and physiological conditions, including disease state. This is particularly true in relation to complex metabolic diseases such as obesity and DM2, characterized by subtle change in expression pattern on many genes and dysregulation of whole biological pathways (see (Barabash, 1963; Merill, 1963; Zagoruiko, 1999) for review). However, selection of an informative disease-related subset of genes which could be used to detect deviation from the normal metabolic status in a quantitative manner presents a significant challenge for this class of diseases. In this paper we introduce two algorithms for feature selection specifically designed to account for the concerted nature of gene activity which underlies metabolic disorders. Algorithm GRAD takes advantage of application of collective deciding rules of "k nearest neighbors" type (kNN). The alternative and complementary algorithm LAD – draws decision from application of logical deciding rules (decision

trees). The performance of our algorithms is demonstrated in a case study of selecting gene expression signatures informative for early diagnosis of DM2.

ALGORITHM GRAD

First effective algorithms for selection of informative subset of dependent features has been proposed by T. Merill and O. Green (Merill, Green 1963). They have described the algorithm of backward deletion (Del). At the same year Ju. Barabash (Barabash, 1963) introduced a forward addition algorithm (Ad). Both algorithms belong to a class of greedy algorithms of unidirectional hill-climbing. On each step algorithm Del excluded the least important feature. Algorithm Ad works in the opposite direction: on each step the subset is coupled with most informative feature. Comparison of different variants of alternating application of Ad and Del algorithms has demonstrated the advantage of combined AdDel algorithm (Zagoruiko, 1999), which works as following: first n1 informative features are selected by Ad algorithm; then n2 of these features (n2 < n1) are discarded by Del. The iterations continue until the desired quality of classification is achieved.

On each step the algorithm produced an optimal solution in a polynomial time. However, it does not guarantee a globally optimal solution. Such solution can be approached by stepwise selection of not just spare features, but "granules", i.e. combinations of a few features. Our experiments demonstrate that if all features are arranged by descending order of individual importance the probability of a particular feature to land in an informative subset quickly drops with its increase of its rank. Thus, most informative combinations of features are likely to be found among features with most individual importance.

The new algorithm GRAD («GRanulated AdDel») developed by our group begins with selection of N<N* most informative features (i.e. power 1 granules) out of original set N*. Out of these features we form all combinations of 2 and 3 features (i.e. power 2 and power 3 granules) and then select N* most informative power 2 and N* most informative power 3 granules. The power order of granules is limited by computational resources and can be scaled up if required by particular research project. On the second step the list of 3N* granules is pipelined into input stream of AdDel algorithm. Starting the algorithm from different elements of this list generates different informative subsystems. Deciding rules of kNN type weight all features in proportion to the number of occurrences in the given subsystem.

ALGORITHM LAD

Logical decision trees are grown "down from the top". Recognizing two patterns we select the feature separating patterns with the least number of mistakes. Than we select the best-separating feature in each group of patterns (objects). The process continues until the desired quality of separation or maximum allowed order of tree branching is reached. This procedure of gradual adding the "best" features is similar to one used in Ad algorithm. Introduction of a complementary deletion technique similar to Del is a logical step in algorithm development.

Resulting algorithm LAD begins with construction of N1 trees of depth 2, where first leafs are N1 most informative features. Among the trees we select one generating the least number of mistakes. Then we fixate the leaf of the tree and construct a set of trees of depth 3, where N2 most informative features are used for the nodes of the next level. We select the tree producing least recognition mistakes on depth 3. The process continues on level 4 and beyond until the algorithm stops.

RESULTS AND DISCUSSION

In this study we used the data previously published by Mootha *et al*. (2003). The data consists of 43 skeletal muscles samples from 43 age-matched male patients divided on three groups: NGT (normoglycemic), DM2 (diabetes mellitus type II) and IGT (insulin-impaired intermediate group). The microarray expression data contains expression measurement for 22365 genes for each sample. Preliminary analysis revealed 5527 genes expressed above the background noise level among at least one of 43 patients. We conducted comparative analysis of algorithms GRAD, AdDel and Exhausting Search (ES) applied for recognition of two classes: NGT (17 patients) and DM2 (18 patients) using the rule kNN (k = 1) in leave-one-out regime. Diagnoses were made with the help of "function of membership" F = 1–2r1/(r1+r2), r1 and r2 – distances from a control point up to the nearest neighbors of the first and second pattern, accordingly. The estimation of informativity of attribute subsystems was made on training set by a method "One-Leave-Out": each of 35 patients was by turns compared to all other patients. Got out on one nearest neighbor of each of two patterns. On distances up to them r1 and r2 value F was calculated. If F > = 1 the decision was accepted in the benefit of the first pattern and if F <0 – for the benefit of the second.

This method of decision-making at use of all of 5527 attributes has correctly distinguished 20 objects from 35. Recognition to 200 attributes having maximal individual informativity, has lowered number of mistakes up to 9. Algorithm GRAD has chosen some tens the subsystems consisting of 3 and 4 attributes on which correct recognition of all of 35 objects turns out.

Averaging the "function of membership" for each patient among different feature subsystems can be used to sort all patients along the line drawn between most and least metabolically fit phenotypes. The results of this ranking of 43 patients on 50 most informative feature subsystems are remarkably consistent with the clinical diagnosis as depicted on Fig. 1.

Here vertical lines separate three classes as they are defined by clinical diagnosis (NGT, IGT and DM2). It is necessary to emphasize, that all samples from the intermediate group are found in the center between NGT and DM2 classes even though they were not employed in the training set.

Using LAD algorithm we have found some trees of depth <4 which distinguish two patterns without mistakes, including one tree of depth 2. Its structure included the following two genes: 220547_s_at (*gb:NM_019054.1 /DEF=Homo sapiens hypothetical protein MGC5560 (MGC5560), mRNA. /FEA=mRNA /GEN=MGC5560 /PROD=hypothetical protein MGC5560 /DB_XREF=gi:12963480 /UG=Hs.233150 hypothetical protein MGC5560 /FL=gb:NM_019054.1*) and 218034_at *(gb:NM_016068.1 /DEF=Homo sapiens CGI-135 protein (LOC51024), mRNA. /FEA=mRNA /GEN=LOC51024 /PROD=CGI-135 protein /DB_XREF=gi:7705631 /UG=Hs.84344 CGI-135 protein /FL=gb:BC003540.1 gb:AF151893.1 gb:NM_016068.1).*

It is necessary to mean, that because of very small quantity of objects the received results have low statistical reliability.

Complexity of the first stage when algorithms choose a subset from N* most individually informative attributes, linearly depends on N. Complexities of formation of granules of capacity 2 and 3 are proportional to a square and a cube of N *, accordingly. Stages of application of algorithm AdDel and algorithm of escalating of a tree have polynomial complexity. Complexity of algorithms depends on quantity M of objects is linearly. Hence, essential restrictions on sizes of data table are not present.

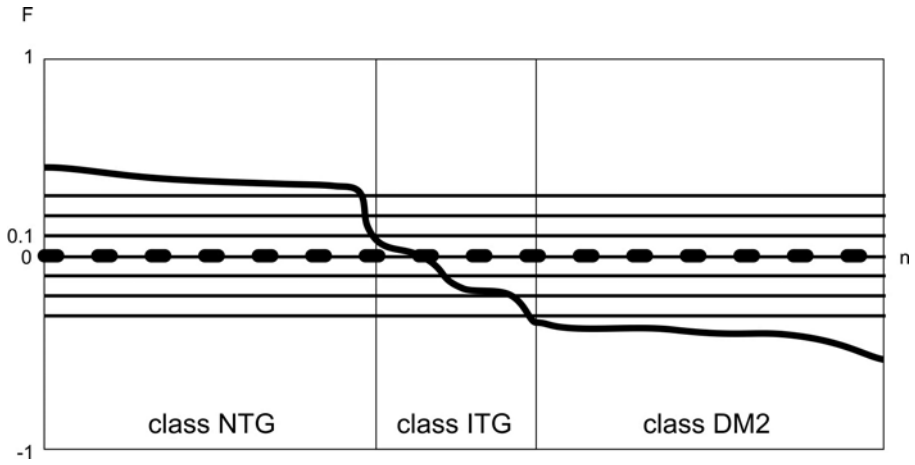


Figure 1. Ranging of patients' microarray phenotypes on the scale from least to most metabolically fit.

CONCLUSION

Algorithms GRAD and LAD are highly effective in selection of subset of informative features in the original data set of extremely high dimensionality. Ranking patients' molecular phenotypes by the unified index can provide important tool for early diagnostics and risk personalized risk estimation of complex multifactorial diseases such as metabolic syndrome and type II diabetes.

ACKNOWLEDGEMENTS

The work was supported by the grant RFBR No. 05-01-00241.

REFERENCES

Barabash Yu. *et al*. (1963) Automatic pattern recognition. Kiev: ed. KVAIU. (In Russ.).

Merill T., Green O. (1963) On the effectiveness of receptors in recognition systems. *IEEE Trans. Inform. Theory*, **IT-9,** 11–17.

Mootha V.K. *et al*. (2003) PGC-1alpharesponsive genes involved in oxidative phosphorylation are coordinately down regulated in human diabetes. *Nat. Genet*., **34**(3), 267–27.

Zagoruiko N.G. (1999) Applied Methods of Data and Knowledge Analysis. Ed. IM SD RAS, Novosibirsk, 270 p.

PART 2. COMPUTATIONAL STRUCTURAL AND FUNCTIONAL PROTEOMICS

CONNP: THE PREDICTION OF THE CONTACT NUMBERS OF THE AMINO ACID RESIDUES IN PROTEINS USING NEURAL NETWORK REGRESSION

Afonnikov D.A. [*1, 2]***, Morozov A.V.*** [1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: ada@boionet.nsc.ru

Key words: protein structure, residue contact numbers, position-specific scoring matrix, neural network

SUMMARY

Motivation: The contact number profile contains important information about residue interactions and, hence, it can be helpful in prediction of the spatial structure of the protein. The residue contact numbers are traditionally interpreted as measure of its solvent exposure.

Results: Here, we propose a neural network regression approach to the prediction of the number of residue contacts of short- and full-range types. Pearson's *r* between the actual and predicted contact numbers varied from 0.531–0.705 for the full-range contacts and they were consistently higher (0.669–0.768) for short-range contacts for all the distance cut-offs at 6,8,10 and 12 Å.

Availability: The program of the contact number prediction CONNP is available at http://wwwmgs2.bionet.nsc.ru/reloaded/.

INTRODUCTION

Residue contact number in a protein is defined by the number of residues at a specified distance around the given residue and closely related to its accessible surface area (Nishikawa, Ooi, 1980; Rodionov *et al*., 1981; Rost, Sander, 1994). Therefore, information about contact numbers can obviously be advantageously used in estimation of residue solvent accessibility in fold recognition problems (Nishikawa, Matsuo, 1993). This information has been also utilized for the prediction of protein contact maps (Fariselli *et al*., 2001). The predictions of contact number rely on the classification and regression approaches. Classification approaches predict the residue state, for example, two state classification with respect to the contact number mean value (Fariselli, Casadio, 2000; Pollastri *et al*., 2002). Regression allows to predict the real value of contact number and proved to be more informative (Kinjo *et al*., 2005; Yuan, 2005; Kinjo, Nishikawa, 2005).

Here, we propose a regression method that uses neural network ensemble for the prediction of contact numbers of both types. Current approach is further development of previously reported work (Afonnikov, 2006) and is extended by using averaging over a set of several neural networks predictions. The algorithm is implemented in the CONNP (CONtact Number Prediction) program available through the Internet.

SYSTEM AND METHODS

Two residues of a protein are in contact, if the distance between their C_α atoms does not trespass the dc threshold (Pollastri *et al*., 2002). In this work we use d_c at the 6, 8, 10, and 12 Å distances. Additionally, for each residue *i*, we partition protein globule. The first part consists of the residues nearby in the polypeptide chain. They define short-range interactions and are separated in the primary structure from the *i*-th residue by not more than 7 positions. The second part is comprised of other residues and defines long-range interactions. Here, we define contact numbers for the full-, long-, and short-range contacts, cnf(i), cnl(i), and cns(i), respectively. It is clear that cnf(i) = cnl(i)+cns(i). Therefore, we predict full- and short-range contact numbers. To train the neural network, high resolution monomeric protein structures from the PDB database (release 109, January, 2005; Berman *et al*., 2000) were extracted. The structures do not contain domains with the same folding type according to the SCOP classification (Andreeva *et al*., 2004). The total number of the full-size monomeric chains that shared no common domains of the same folding type was 339. The chains were divided into 3 samples (a, b, and c), consisting of 113 chains each, with approximately the same number of residues. The neural networks were trained, tested and validated on this abc dataset. For the additional test, we selected those protein structures that were supplemented to the PDB database in 2005, i.e. later than the proteins included in the training sample. We chose the fully resolved structures with a value of pairwise sequence similarity < 25 % between each other and with those of the abc dataset; 408 sequences in all (2005 dataset).

ALGORITHM

To predict contact numbers, a set of fed-forward neural networks of the first (NN1) and the second (NN2) levels was used. Networks at the same level had input data of the same type. All the networks had a topology with a single hidden layer and sigmoidal normalized exponential transfer functions. Fig. 1*a* shows the structure of the NN1. The NN1 predicts the contact number on the basis of the PSSM data. The PSSM matrix is built on the basis of the PSI-BLAST multiple alignment program, 3 iterations with default parameters (Altschul *et al*., 1997). The NN1 comprised a single hidden layer and a single neuron at the output layer. To optimize the NN1 parameters, the mean absolute deviation error (MAE) was used: $MAE = (1/N)\cdot\Sigma|cn0(i) - cn(i)|$, where N is the test sample size, cn0(i) is the observed contact number, and cn(i) is the predicted contact number for the residue i. We used the abc dataset described above for the training/test/validation cycles, i.e. every one of the samples was trained/tested/validated in different combinations with other two samples. Four NN1 networks were built for each combination. The networks had a hidden layer with 20, 40, 60, and 80 neurons; then, the output of 4 networks were arithmetically averaged. The average was the result of the NN1 for each data combination.

The NN2 was arranged as shown in Fig. 1*b*. The input data resulted from the NN1. For the *i*-th residue, a sliding window of 41 positions in length was considered. Predictions were made for the central residue in the window. The input data vector was composed of 41 contact number values predicted by the NN1. For the N- and C- residues, the parameter values in the sequence range were taken as -1. All the NN2 contained a single hidden layer with 20 neurons. Like in the NN1, there was a single value at the output, namely the contact number value for the *i*-th residue. The training procedure was the same as for the NN1 networks.

Fig. 1*c* is a general layout of the neural network ensemble for the prediction of the real values of the contact number. The parameters for the NN1 with different neuron numbers in the hidden layer were obtained for the training/test/validation data sets (abc, acb, bac, bca, cab, cba), $4 \times 6 = 24$ networks of the NN1 were built. Six NN2 networks were built

for each data set combination. The final prediction was the averaging of the predictions for 6 NN2 and rounding of the value to the nearest integer. Thus, the structure of the predictor ensemble architecture is comprised of 30 neural networks for each contact distance and contact type. The total number of built predictors was 8.

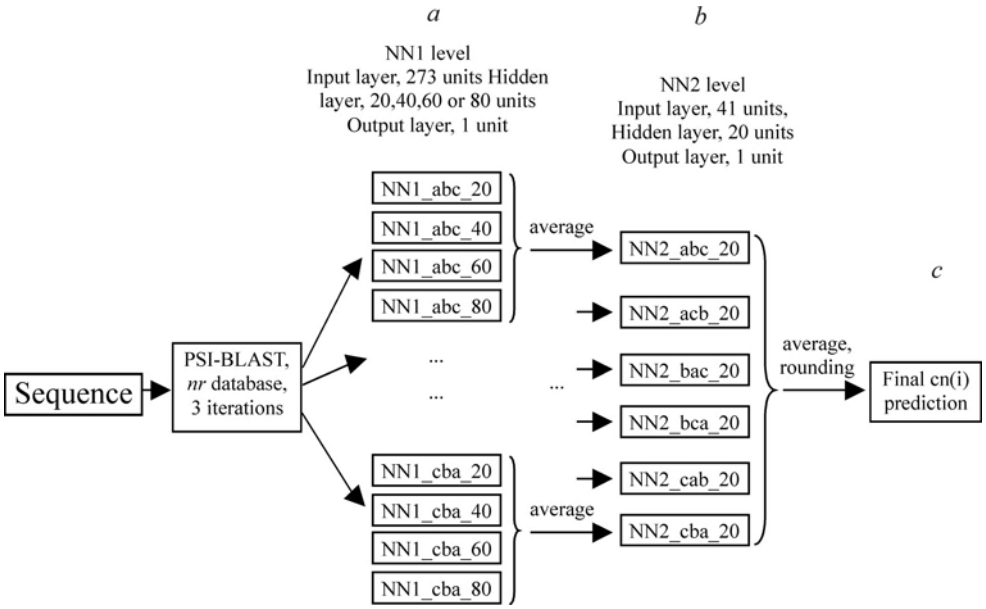


Figure 1. The neural network ensemble for the prediction of the real values of the contact numbers: *a* – the neural network at the first level (NN1) uses PSSM and Information weights yielded by the BLAST program as input data; *b* – the neural network at the second level (NN2) uses the NN1 predictions as input parameters; *c* – the overall ensemble structure that uses the averaged predictions of the neural networks (the designation of the neural network NNL_xyz_K in the figure corresponds to the network at the L level, to the xyz set of the training/test/validation data, and the number of K units in the hidden layer of the network). For detailed description, see text.

RESULTS AND DISCUSSION

The accuracy for the contact number prediction was estimated on the 2005 dataset (Table 1). It is evident that on average, short-range contact numbers predicted with higher accuracy compared with full-range contact numbers. For the full-range contact numbers the performance of our program is comparable with that reported by SVM regression method for the prediction of the Cβ-Cβ distance discrete contact number (r = 0.64, 0.66, and 0.69 for d_c = 8, 10, and 12 Å, respectively, in the case when the PSSM matrix was used as input data) (Yuan, 2005). We also performed the residue classification based on the contact number predicted by the current method and estimated the performance of classification approach as the fraction of residues classified correctly Q_2. For the contact distances of 6, 8, 10, and 12 Å we obtained Q_2 values 0.726, 0.736, 0.753 and 0.766 respectively. These values slightly outperforms previously reported 0.7324 (6 Å), 0.7095 (8 Å), 0.7213 (10 Å) and 0.7409 (12 Å) (Pollastri *et al.*, 2002).

Thus, the CONNP program can be used in the prediction of contact numbers as additional source of information and can be applied with the results of other methods to obtain better contact number estimates.

Table 1. CONNP Performance on the 2005 dataset. The contact number of type t and distance d will be denoted cntd

Contact number	MAE[a]	r[b]	*DevA*[c]
cn_{f6}	1.073	0.531	0.848
cn_{f8}	1.750	0.628	0.780
cn_{f10}	3.350	0.667	0.750
cn_{f12}	5.578	0.705	0.720
cn_{s6}	0.891	0.718	0.698
cn_{s8}	0.995	0.669	0.746
cn_{s10}	1.288	0.713	0.702
cn_{s12}	1.242	0.768	0.641

[a] Mean absolute error. [b] Pearson's r. [c] Absolute deviation (Kinjo *et al*., 2005).

ACKNOWLEDGEMENTS

The authors are grateful to Lokhova I.V. for technical assistance. The work is supported by the U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF) within the Basic Research and Higher Education Program (Y1-B-08-20), the Ministry of Education of the Russian Federation grant РНП.2.1.1.4935, Russian Foundation of the Basic Research (05-04-49141-a, 05-07-98012-p), SB RAS integration projects 49 and 115, Innovation project of Federal Agency of Science and innovation IT-CP.5/001.

REFERENCES

Afonnikov D.A. (2006) Prediction of contact numbers of amino acid residues using a neural network model. In Kolchanov N., Hofestaedt R., (eds), *Bioinformatics of Genome Regulation and Structure II*. Springer Science+Business Media, Inc. 2006, pp. 297–304.

Altschul S.F., Madden T.L., Schaffer AA., 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.

Andreeva A., Howorth D., Brenner S.E., Hubbard T.J., Chothia C., Murzin A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucl. Acids Res*., **32**, D226–D229.

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.

Fariselli P., Casadio R. (2000) Prediction of the number of residue contacts in proteins. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **8**, 146–151.

Fariselli P., Olmea O., Valencia A., Casadio R. (2001) Progress in predicting inter-residue contacts of proteins with neural networks and correlated mutations, *Proteins*, **45** (Suppl 5), 157–162.

Kinjo A.R., Horimoto K., Nishikawa K. (2005) Predicting absolute contact numbers of native protein structure from amino acid sequence. *Proteins*, **58**, 158–165.

Kinjo A.R., Nishikawa K. (2005) Predicting secondary structures, contact numbers, and residue-wise contact orders of native protein structures from amino acid sequences using critical random networks. *Biophysics*, **1**, 67–74.

Nishikawa K., Ooi T. (1980) Prediction of the surface-interior diagram of globular proteins by an empirical method. *Int. J. Pept. Protein. Res.*, **16**, 19–32.

Nishikawa K., Matsuo Y. (1993) Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies. *Protein Eng.*, **6**, 811–820.

Pollastri G., Baldi P., Fariselli P., Casadio R. (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.

Rodionov M.A., Galaktionov S.G., Akhrem A.A. (1981) Prediction of the degree of exposure of amino acid residues in globular proteins. *Dokl. Akad. Nauk SSSR*, **261**, 756–759.

Rost B., Sander C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.

Yuan Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, **6**, 248.

A TOOL FOR COMPARATIVE ANALYSIS OF SOLVENT MOLECULES IN PDB STRUCTURES

Aksianov E. [*1,2]***, Zanegina O.***[3]***, Alexeevski A.***[2]***, Karyagina A.***[4,5]***, Spirin S.***[2]

[1] Virology department, Biological faculty, Moscow State University, Moscow, Russia; [2] Belozersky Institute, Moscow State University, Moscow, Russia; [3] Bioengineering and Bioinformatics faculty, Moscow State University, Moscow, Russia; [4] N.F. Gamaleya Research Institute of Epidemiology and Microbiology, Moscow, Russia; [5] Institute of Agricultural Biotechnology, Moscow, Russia

[*] Corresponding author: e-mail: evaksianov@belozersky.msu.ru

Key words: structural water molecule, hydrogen bond, comparative analysis of related structures

SUMMARY

Motivation: Water molecules immobilized on the protein or DNA surface are known to play an important role in the inter-molecular contacts, protein-protein or protein-DNA complexes stabilization. Comparative analysis of related 3D structures allows to predict locations of such water molecules on intermolecular interface.

Results: We have developed and implemented the algorithm **wLake** detecting "conserved" water molecules i.e. those located in almost the same positions in a set of superimposed structures of related proteins or macromolecular complexes. The water molecules of different complexes are represented as vertexes of a certain graph and the conserved molecules correspond to maximal cliques in the graph. **wLake** was used to analyze water molecules in several structurally characterized protein families. Despite of exponential algorithm complexity, the program works appropriately fast for dozens of superimposed structures.

Availability: http://monkey.belozersky.msu.ru/.

INTRODUCTION

Water molecules immobilized on a protein surface or on an interface of two macromolecules play an important role in intermolecular (protein-protein, protein-DNA etc.) interactions and macromolecular complexes stabilization. Protein Data Bank entries obtained by X-ray diffraction with resolution better than 2.5 Å typically report the coordinates of a number of water molecules. The availability of 3D structures of homologous proteins provides a possibility to select those water molecules that are located in the same positions and form the same hydrogen bonds. It could be supposed that those molecules, referred as structural water molecules (SWMs), correspond to hydration sites on protein or nucleic acid surfaces in solution. This approach was explored in a growing number of studies devoted to protein-DNA interactions (see, for example, Karyagina *et al*., 2005), formation of protein oligomers (Bella *et al*., 1995) and protein-substrate complexes (Ogata, Wodak, 2002).

We present here an automatic tool (named **wLake**) finding structural water molecules in superimposed 3D structures; unlike other analogous tools, e.g. Sanschagrin, Kuhn (1998), our tool is augmented with a module estimating the reliability of obtained results.

METHODS AND ALGORITHMS

Definitions. Two water molecules from different structures are called *close* if the distance between the centers of their oxygen atoms in superimposed structures is less than a threshold and (optionally) they form hydrogen bonds with the same amino acid residues and/or nucleotides. Default threshold value is 1.5 Å. A set of water molecules from different structures is called *a cluster of water molecules* if each two molecules of the set are close.

Input data for identification of SWMs are superimposed 3D structures of related proteins or macromolecular complexes. Atoms of all structures are assumed to be immersed in the same coordinate frame.

Algorithm. The goal of the algorithm is to find all maximal (i.e. non-extendable) clusters of more than a given number `Min` water molecules. At the first stage, we construct a graph representing the set of water molecules. A graph vertex corresponds to a single water molecule from any structure. Two vertices are connected by an edge if the respective water molecules are close and therefore belong to different structures. Thus, maximal clusters of waters correspond to maximal cliques in the graph (a clique is a subset of vertices such that any pair of vertices is connected by an edge). A simple algorithm is used to find all maximal cliques of more than `Min` vertices (see **wLake** homepage for details). Despite it has an exponential complexity, the algorithm demonstrated a reasonable efficiency for several dozens of superimposed structures.

Parameters of the algorithm are (1) the distance threshold for regarding water molecules as close (1.5 Å by default), (2) the minimal number `Min` of water molecules in a cluster (3 by default), (3) the distance threshold for a hydrogen bond length (3.5 Å by default).

The algorithm was implemented as the **wLake** program available via Web interface. The preliminary version of the Web service can be found at http://monkey.belozersky.msu.ru/.

The result of the program is a list of clusters of water molecules. To estimate the reliability of each detected cluster, we developed a special program **WLStat**. This program 100 times repeats the following two-step procedure. First, a "randomized PDB file" is generated. The file has PDB format and contains (1) the superimposed macromolecules from the input data and (2) water molecules with randomized coordinates; the number of water molecules in each of superimposed structures and the distribution of distances between water molecules and macromolecules are the same as in input data. Second, clusters of water molecules in the randomized PDB file are detected by **wLake**. This allows to approximate the distribution of cluster sizes and to assign an E-value to each detected cluster. Clusters with small E-value ($E < 0.01$ by default) are considered as clusters of SWMs.

wLake was used to detect clusters of SWMs in several protein families (constant domains of TCR, N-terminal domains of NFκB, transketolases and others). The lists of family representatives were obtained from SCOP database (Murzin *et al.*, 1995). SSM server (Krissinel *et al.*, 2004) was used to superimpose the structures. All detected clusters of SWMs were analyzed manually using RasMol program (Sayle, Milner-White, 1995).

IMPLEMENTATION AND RESULTS

wLake was tested on structures of homeodomain-DNA complexes. In the work of Karyagina *et al.*, 2005, we have reported 8 hypothetically functionally important clusters of water molecules found in protein-DNA interfaces of 32 superimposed X-ray structures from different homeodomain subfamilies. Seven of these clusters contain water molecules of 11 or more structures. **wLake** program with the distance threshold 1.5Å, the minimal cluster size 10, and the E-value threshold 0.01 has detected 12 clusters of structural water

molecules on the protein-DNA interface of same complexes. Five of them coincide with those reported in (Karyagina *et al*., 2005). The other two clusters from (Karyagina *et al*., 2005) can be easily extracted from the remaining seven clusters detected by **wLake.** Namely, the seven **wLake** clusters can be divided in two groups; clusters from each group have essential mutual intersection and no intersections with other clusters. The set of molecules belonging to each group coincide with one of earlier reported clusters of water molecules. We conclude, that **wLake** is appropriate for SWM detections.

To compare our program with another water cluster detector **WatCH** (Sanschagrin, Kuhn, 1998), we tested the latter on the same set of homeodomain-DNA complexes. **WatCH** with the same distance threshold 1.5 Å detected 7 clusters on protein-DNA interface containing 10 or more water molecules. One **WatCH** cluster contains more than one water molecule from the same structure, which is allowed by its algorithm but seems to be in conflict with the purpose to find SWMs. Five clusters coincided with those reported in the work Karyagina *et al*. (2005). The remaining cluster is a part of a cluster reported in the cited work and detected by **wLake**.

Analysis of differences in the results of two programs showed that the clustering algorithm of **WatCH** can divide big and compact clusters of waters into smaller parts, which is not the case of **wLake**. Probably, the biggest clusters of water molecules are more adequate solutions of the problem. **wLake** detects maximal but sometimes intersecting clusters of SWMs. Intersecting clusters should be considered by an expert. and merged if necessary. To facilitate the analysis, **wLake** generates a RasMol script for visualization detected clusters of SWMs.

We have also implemented **wLake** to analyze immunoglobulin-like beta-sandwich domain of *Escherichia coli* thiol-disulfide interchange protein. One of its strand-like segments in a beta sheet is not connected with the neighboring strand with hydrogen bonds. To explain the stability of this segment, we have detected SWMs in 6 X-ray structures of the protein. Using **wLake,** 6 clusters of SWMs forming water bridges between the segment and the neighboring beta-strand were detected. The found SWMs can explain the unusual for the immunoglobulin fold feature (Fig. 1). **wLake** program was used to detect SWMs in structures of transketolases from *S. cerevisiae*, *E. coli*, *Zea mayse* and *Leishmania mexicana*. Eight structures (two subunits from each homodimer) were superimposed. We have detected 27 clusters of SWMs containing 8 water molecules (one molecule from every structure) were detected. The SWMs will be taken into account in homology modeling of the human transketolase TKTL1.

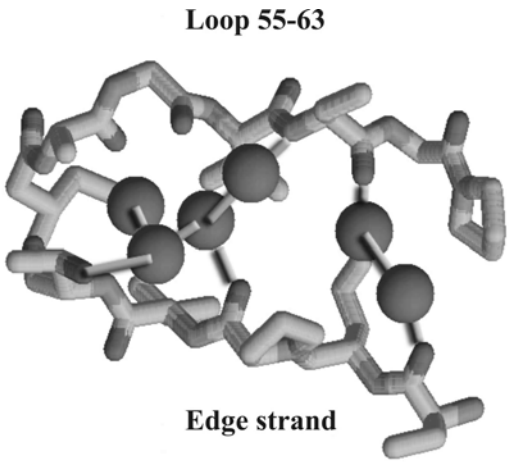


Figure 1. Water-mediated contacts between the loop 55–63 and the beta-strand 80 87 in 1L6P.

DISCUSSION

It is a common observation that many water molecules reported in X-ray solved structures are not hydrogen bonded to proteins, nucleic acids, ligands or other water molecules. The majority of such molecules seems be artifacts of the method; the same could be true even for some water molecules bound with protein and/or nucleic acids. We have postulated that water molecules, bound with protein and/or nucleic acid and detected practically at the same position in several related structures, SWMs, are more reliable than other water molecules; they can be functionally important. For example, they can play an important role in intermolecular interaction. We believe that those molecules correspond to structural water molecules in solution, which occupy hydration sites for a relatively long time. Interfacial SWMs should be taken into account in studies of protein-DNA and protein-ligand interactions, and in drug-design studies.

wLake, a special tool for automatic detection of SWM clusters in a set of related superimposed 3D structures, was developed and tested. It is equipped with a module **WLStat** which estimates the reliability of SWM clusters. These tools facilitate and standardize the procedure of SWM finding.

ACKNOWLEDGEMENTS

This work is partially supported by Ludwig Institute for Cancer Research (CRDF GAP grant RB0-12771) and RFBR grants Nos 06-04-49558 and 06-07-89143.

REFERENCES

Bella J. *et al*. (1995) Hydration structure of a collagen peptide. *Structure*, **3,** 893–906.

Berman H.M. *et al*. (2000) The Protein Data Bank. *Nucl. Acids Res.*, **28,** 235–242.

Karyagina A. *et al*. (2005) The role of water in homeodomain-DNA interaction. In Kolchanov N. and Hofestaedt R. (eds), *Bioinformatics of Genome Regulation and Structure II.*, Springer Science+Business Media. pp. 247–257.

Krissinel E. (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Cryst.* **D60**, 2256–2268.

Murzin A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*. **247,** 536–540.

Ogata K., Wodak S. (2002) Conserved water molecules in MHC class-I molecules and their putative structural and functional roles. *Protein Engeneering*, **15**, 697–705.

Sanschagrin P., Kuhn L. (1998) Cluster analysis of consensus water sites in thrombin and trypsin shows conservation between serine proteases and contributions to ligand specificity. *Protein Science*, **7**, 2054–2064.

Sayle R., Milner-White E. (1995) RasMol: Biomolecular graphics for all. *Trends in Biochem. Sci.*, **20**, 374.

ANALYSIS OF THE TERTIARY STRUCTURE OF THE PPAR AND RXR TRANSCRIPTIONAL FACTORS AND THEIR MUTANT VARIANTS

Aman E.E.*[*], *Demenkov P.S.*, *Ivanisenko V.A.
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: aman@bionet.nsc.ru

Key words: protein tertiary structure, protein functional sites, site recognition, thermodynamic stability

SUMMARY

Motivation: Analysis of the tertiary structure of proteins is a key approach to the study of their functions. An important aspect in the detection of novel regulatory mechanisms in the gene networks is the recognition of the protein-protein interaction sites in the spatial structures of the transcriptional factors and analysis of the effects of mutations on the physical properties of their molecules.

Result: The tertiary structure of the PPAR and RXR transcriptional factors were studied using the method of functional site recognition and also the method estimating the effect of mutations on the thermodynamic stability of protein structure. The potential interactions of the PPARs and RXRs molecules with other proteins were detected using the PDBSiteScan method. To estimate the effect of mutations associated with the development of human diseases on the thermodynamic stability of the PPARs tertiary structure, we applied the modified KRAB method.

INTRODUCTION

The PPAR and RXR transcriptional factors play an important role in the regulation of energy metabolism in humans and animals. Forming a heterodimer, they bind to the sites on DNA in the promoter gene regions and regulate their expression. The activity of the PPAR/RXR heterodimer is modulated by the interaction of its components with various molecules: low molecular weight ligands (fatty acids, eicosanoids, fibrates, to name a few), and also proteins: coactivators and corepressors. The PPAR/RXR heterodimer is also subject to phosphorylation. Search and detection of new PPAR and RXR binding sites to other proteins would hopefully disclose novel regulatory processes of energy metabolism. There are data in the literature indicating that physiological disorders are associated with point mutations in the PPARs genes. Such mutations not only disrupt the structure of the functional protein domains, they can destabilize a protein within the cell. Analysis of changes in the thermodynamic stability of PPARs molecule upon the rise of point mutation would give clues to the possible causes of mutation-associated diseases.

METHODS AND ALGORITHMS

The tertiary structures of the PPARs and RXRs factors were retrieved from the PDB database. The full-size sequences of the PPARs and RXRs are 468 (PPARα), 441

(PPARδ), 505 (PPARγ), 462 (RXRα) and 533 (RXRβ) amino acids. The tertiary structures stored in the PDB database contain only fragments of full-size sequences corresponding to the ligand-binding domain. For analysis, we chose 10 PPAR and RXR human structures (PDB ID: 1i7g, 1k7l (PPARα); 1y0s, 1gwx (PPARβ(δ)); 1i7i, 1wm0; 1fm6 (PPARγ); 1dkf (RXRα); 1h9u, 1uhl (RXRβ)).

The tertiary PPAR and RXR structures were tested for the presence of potential sites for protein-protein interactions using the PDBSiteScan algorithm (Ivanisenko *et al*., 2004). The algorithm is based on the alignment of the examined tertiary protein structures with known binding sites collected in the PDBSite database (Ivanisenko *et al*., 2005a). Our search was focused only on the sites for protein interactions whose maximum deviation of protein structure from the known site structure (max. dist) was not greater than 4 Å. The models for the protein-protein interactions were visualized using the ViewerLite program that makes possible the representation of text descriptions of the spatial structures of the molecules as 3d models.

For analysis of the effect of point mutations on the thermodynamic stability of the PPARs tertiary structure, we utilized a modification of the KRAB method (Ivanisenko, 2005b). Under study were the following mutations: Val227Ala (PPARα) (Yamakawa-Kobayashi *et al*., 2002), Pro467Leu, Val290Met (Barroso *et al*., 1999), Phe388Leu (Hegele *et al*., 2002), Arg425Cys (PPARγ) (Agarwal, Garg, 2002).

RESULTS AND DISCUSSION

Potential sites for protein-protein interactions in the PPARs and RXRs molecules. Using the PDBSiteScan program, we detected sites for protein-protein interactions and built molecular complexes for three human PPAR isotypes and two RXR isotypes. In the case when the maximum deviation was less than 4Å, the number of predicted proteins was 51 for PPARα, 40 for PPARβ, 46 for PPARγ**,** 49 for RXRα and 37 for RXRβ**.** Three interactions (RXRα – annexin II, RXRα – tyrosine-kinase p56-lck, PPARα – histone H4) of particular interest are depicted in Fig. 1. Annexin II is a calcium-dependent phospholipid-binding protein of the cell membrane (Deora *et al*., 2004). What may be the biological significance of the RXRα – annexin II interaction we predicted (Fig. 1*a*)? The question remains open. Answers might help to establish new functions of the RXRα transcriptional factor.

The potential capacity of the RXRα – tyrosine-kinase p56-lck complex (Fig. 1*b*) is consistent with the data indicating that the RXRα molecule is subject to phosphorylation. It is now known that the modulator RXRα domain contains 3 sites for the phosphorylation by MAP kinases. The phosphorylation degree may affect transcriptional activity of RXRα, and also its stability in the cell (Gianni *et al*., 2003).

The detected potential capacity of the protein-protein PPARα – histone H4 interaction (Fig. 1*c*) suggests the existence of the following regulatory mechanism for gene expression: 1) In the DNA with loose nucleotide packing, histone H4 molecules can bind with low affinity to PPARα resulting in PPARs accumulation nearby DNA; 2) the low affinity PPARα – histone H4 interaction allows PPARα in complex with RXR to move along DNA as a result of skipping over from one histone to another that hastens the recognition of the binding sites on DNA. The mechanism appears feasible when taking into account the fact that the regions that contain PPARα binding sites have a high potential for nucleosome formation (Levitsky *et al*., 2004). We further intend to analyze in more detail the PPARα – histone H4 binding mechanism relying on modeling of molecular dynamics.

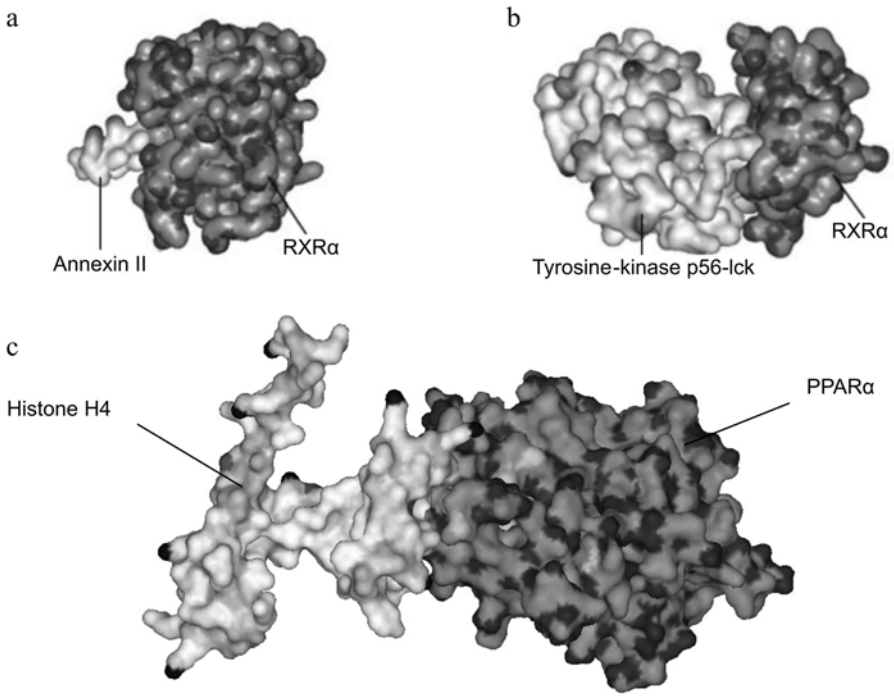


Figure 1. The potential protein-protein complexes. *a* – the RXRα – annexin II complex; *b* – the RXRα – tyrosine-kinase p56-lck complex; *c* – the PPARα – histone H4 complex. The PPARα and RXRα is a dark surface; annexin II, histone H4, tyrosine-kinase p56 lck is a light surface.

Effect of mutations on the stability of the PPAR molecules. The effect of the mutations on the thermodynamic stability of the PPARs tertiary structure was studied using the modified KRAB method (Ivanisenko *et al*., 2005b). It was found that mutations at position 227 in PPARα and at positions 290, 388, and 425 in PPARγ produce a decrease in thermodynamic stability. Mutations in the PPARγ molecule at position 467 are without effect on thermodynamic stability.

Table 1. Physiological disorders in humans associated with point mutations in the PPARs genes. Effect of point mutations on the thermodynamic stability of PPAR molecules

Gene	Mutation	Effect of mutation at the molecular level	Population effect of mutations	Mutation effect on thermodynamic stability
PPARα	Val227Ala	Mutation between DNA binding and ligand-binding domain	Carriers of the Ala227 allele had a lower total cholesterol level in plasma. A particularly strong association was established for women older 45.	Decrease in TS
PPARγ	Pro467Leu, Val290Met	Mutation in the ligand binding domain	Not sensitive to insulin, type 2 diabetes, hypertension.	Mutation at: 290 decrease in TS 467 don't affect TS
PPARγ	Phe388Leu	Mutation in the eighth helix of the ligand-binding domain.	Lipodistrophy (hereditary partial lipodistrophy)	Decrease in TS
PPARγ	Arg425Cys	Mutation in the loop between the helixes 9 and 10 of the ligand-binding domain.	Partial hereditary lipodistrophy.	Decrease in TS

The current results indicate that the mutations described in the literature reduce the stability of the PPAR molecules and, accordingly, accelerate PPARs decay in the proteosomes. Possibly, increase in the decay rate of the PPAR molecules, along with impairment of the structures of their functional domains, contribute to the development of the above listed human diseases.

ACKNOWLEDGEMENTS

Work was supported in part by Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49283, 06-04-49556), the CRDF Rup2-2629-NO-04, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction" federal goal-oriented technical program "Study and development of priority directions in science and technique, 2002–2006", Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science".

REFERENCES

Agarwal A.K., Garg A. (2002) A novel heterozygous mutation in peroxisome proliferator-activated receptor-gamma gene in a patient with familial partial lipodystrophy. *J. Clin Endocrinol. Metab*, **87,** 408–411.

Barroso I., Gurnell M., Crowley V.E., Agostini M., Schwabe J.W., Soos M.A., Maslen G.L., Williams T.D., Lewis H., Schafer A.J., Chatterjee V.K., O'Rahilly S. (1999) Dominant negative mutations in human PPARgamma associated with severe insulin resistance, diabetes mellitus and hypertension. *Nature*, **402,** 880–883.

Deora A.B., Kreitzer G., Jacovina A.T., Hajjar K.A. (2004) An annexin 2 phosphorylation switch mediates p11-dependent translocation of annexin 2 to the cell surface. *J. Biol. Chem*., **279,** 43411–43418.

Gianni M., Tarrade A., Nigro E.A., Garattini E., Rochette-Egly C. (2003) The AF-1 and AF-2 domains of RAR gamma 2 and RXR alpha cooperate for triggering the transactivation and the degradation of RAR gamma 2/RXR alpha heterodimers. *J. Biol. Chem.*, **278,** 34458–34466.

Hegele R.A., Cao H., Frankowski C., Mathews S.T., Leff T. (2002) PPARG F388L, a transactivation-deficient mutant, in familial partial lipodystrophy. *Diabetes*, **51**, 3586–3590.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucl. Acids Res*., **32** (Web Server issue), 549–554.

Ivanisenko V.A., Pintus S.S., Grigorovich D.A., Kolchanov N.A. (2005a) PDBSite: a database of the 3D structure of protein functional sites. *Nucl. Acids Res*., **33**(Database issue), 183–187.

Ivanisenko V.A., Pintus S.S., Demenkov P.S., Krestyanova M.A., Litvenko E.K., Grigorovich D.A., Debelov V.A. (2005b) FASTPROT: a computational workbench for recognition of the structural and functional determinants in protein tertiary structures. In *Bioinformatics of Genome Regulation and Structure II*. Springer Science+Business Media, Inc. 2005, 305–316

Levitsky V.G., Ignatieva E.V., Busygina T.V., Merkulova T.I. (2004) Analysis of the context features of SF-1 binding site and development of a criterion for SF-1 regulated gene recognition by the SiteGA method, *BGRS-2004*, Novosibirsk, **1,** 119–122.

Yamakawa-Kobayashi K., Ishiguro H., Arinami T., Miyazaki R., Hamaguchi H.A. (2002) Val227Ala polymorphism in the peroxisome proliferator activated receptor alpha (PPARalpha) gene is associated with variations in serum lipid levels. *J. Med. Genet.*, **39**, 189–191.

MOLECULAR MODELING OF *B. CEREUS* HEMOLYSIN II, A PORE-FORMING PROTEIN

Bakulina A.Yu.*[*1], Sineva E.V.[2], Solonin A.S.[2], Maksyutov A.Z.[1]
[1] State Research Center of Virology and Biotechnology "Vector", Koltsovo, Novosibirsk region, 630559, Russia; [2] Institute of Biochemistry and Physiology of Microorganisms, RAS, Pushchino, Moscow Region, 142290, Russia
[*] Corresponding author: e-mail: nastya@ns.uic.nsu.ru

Key words: HlyII, ionic channels, homology modeling, de novo modeling, docking

SUMMARY

Motivation: Hemolysin II (HlyII) is one of several cytolytic proteins produced by Bacillus cereus, an opportunistic human pathogen. This toxin is secreted in a soluble form and cause its cytolytic effect by assembling into transmembrane pores. HlyII is proposed to be an important factor of B.cereus pathogenity and adaptivity.

Results: We have built the homology model of the HlyII heptameric ionic channel. The model lacks the C-terminal domain of HlyII. Using de novo modeling, we have proposed a possible structure of the C-terminal domain. Alternative 6-meric and 8-meric structures of HlyII ionic channel were built with symmetric multimer docking technique. Analyses of the models helped to explane some experimental results and plan further experimental work.

INTRODUCTION

Bacillus cereus is an opportunistic human pathogen well known as a food and a cosmetic pollutant and was found to be infectious agent of endophtalmitis, food poisoning sometimes lethal and wound defeats (Granum, 1997). Closely related microorganisms of B.cereus group (*B. cereus*, *B. thuriengiensis* and *B. anthracis*) produce wide spectrum of toxins and demonstrate distinct pathogenic properties. The mechanisms, including pathogenicity, that make these bacteria so adaptive to the different ecological niches are not fully understood.

One of several cytolytic proteins produced by B. cereus is hemolysin II (HlyII). Based on its amino acid sequence, HlyII is a member of the family of oligomeric β-barrel channel-forming toxins (β-PFT). The most studied representative of the family is S. aureus α-hemolysin (αHL), which is major factor of pathogenicity of the microorganism. Its cytotoxic, biochemical and pore-forming properties are studied in the deep details and mediated by the formation of heptameric channels. The primary mechanism of cell damage is leakage of ions, water and small molecular weight molecules. The high resolution X-ray structure of the αHL heptameric assembly was determined (Song *et al*., 1996). Properties of HlyII and αHL ionic channels are quite differ (Miles *et al*., 2002).

METHODS AND ALGORITHMS

We used MUSCLE (Edgar, 2004) for multiple alignment of all homological sequences found by BLAST and determined highly conserved residues before homology modeling.

The only appropriate template for the HlyII heptamer in PDB databank is the structure of the α-hemolysin from S. aureus, 7AHL (Song *et al*., 1996). About 100 residues at C-terminal part of HlyII are absent in 7AHL and cannot be modeled using direct homology modeling approach. Several alternative alignments of HlyII without C-term vs. 7AHL were made. Frankenstein server (Kosinski *et al*., 2005) was used to generate the optimal model of the heptamer.

The C-terminal domain of HlyII was modeled de novo with Rosetta (Bonneau, 2001) program, because fold prediction metaserver GeneSilico Metaserver (Kurowski, 2003) showed no suitable template for this domain. About 5 % of 5000 generated decoys were clustered in the biggest cluster, which corresponds to moderate confidence level. The search of 3D-structural homologs for the de novo modeled C-terminal domain was performed using FATCAT server (Ye *et al*., 2004). The structure of *E. coli* CyaY protein, 1EW4 (Cho *et al*., 2000), belonged to the frataxin family, was found appropriative to build the homology model with Modeller8v0 (Sali *et al*., 1993).

Hexameric and octameric forms of ionic channel were built with M-ZDOCK (Pierce et al, 2005) program of symmetrical protein docking. The monomeric structure was taken from the previously built heptameric model. Stem domain and the other part of HlyII were docked separately, and their spatial superposition was used as a template for homology modeling with Modeller8v0 (Sali *et al*., 1993).

IMPLEMENTATION AND RESULTS

The HlyII protein sequence (excluding 94 C-terminal amino acids) has enough degree of similarity with sequence of αHL for homology modeling (31 % identity, 60 % positives). The modelled heptameric complex of HlyII has a mushroom shape and comprises the cap domain, the stem domain and seven rim domains. The oligomer measures about 10 nm in diameter, 10 nm in height and pore diameter is varied from 1 to 4 nm along the channel length. The diameter of the heptameric HlyII pore is around 2 nm at the entrance, maximal – 4 nm inside the cap, 1.2 nm – at the beginning of the transmembrane stem and from 1 nm to 1.6 nm along the rest of the stem. Narrowest diameters highly depend on positions of long sidechains of charged residues inside the pore and, as a result, can vary at different conditions.

The stem region of HlyII contains more charged amino acid groups inside than the same region of αHL. There are four belts of charged residues on the interior surface of the stem domain of HlyII, and only two belts in the corresponded surface of αHL. In contrast to αHL, there are no solvent-exposed hydrophobic residues on the interior surface of the pore. Also, αHL has seven small side holes in the stem, they are connected by H144 residues surrounding the stem and proposed to comprise the pH sensor that gates conductance of the αHL. The corresponding residue in the HlyII structure is E141 and the small holes are preserved in the model structure.

Residues involved in monomer-monomer interactions are not conserved well in the primary protein structures of two hemolysins. Probably, predominantly backbone-backbone type of interactions between the subunits in the pore, as well as absence of strict geometric limitations for this type of interactions lead to the spreading of neutral amino acids substitutions after gene divergence. In contrast, part of the HlyII rim domain (176–197, 250–259 aa), probably interacting with membrane, has the highest identity with the template sequence.

The diameters of the stem domains of HlyII 6-, 7- and 8-mers measured as Cα-Cα distances are 2, 2.6 and 3.2 nm, respectively. The models of 6- and 8-mers have no clashs or other structural problems.

The C-terminal domain most likely lies outside the pore. For this domain we propose α+β frataxin-like fold. We did not make any assumptions about the possible function of the C-terminal domain and used only the aminoacid sequence and *de novo* models for the template choise.

DISCUSSION

The overall structure of HlyII 7-mer is very similar to the structure of αHL, but some important distinctions were found. It can explane known differences in properties of HlyII ionic channel in compare with αHL one.

It is still have to be proved if HlyII can really form heterogenious channels, but some experimental observations support this theory. Using these models, it is possible to estimate single-channel conductances of 6- and 8-meric channels and compare it with experimental results.

Frataxin-like fold for the C-terminal domain seems to agree with experimental data, since expression of HlyII is regulated by iron. This type of regulation is typical for proteins which involved in iron metabolism. Frataxin proteins play important role in the iron uptake in both prokaryotes and eukaryotes by binding and retaining iron ions with use of negatively charged surface islands. Existence of iron-binding properties of the C-terminal domain is questionable, because despite of possible fold similarity, the patches of negatively charged amino acids on the protein surface are not pronounced, and acid residues are seem to be replaced by neutral or positive charged ones. However, the possibility that the small negatively charged islands, found at the C-terminus, can be a part of bacterial iron uptake system should be checked experimentally.

HlyII is not only a drug target, but also a perspective scaffold for the design of novel ionic channels with desirable properties. HlyII modeling can help to understand the fundamental properties of ionic channels.

ACKNOWLEDGEMENTS

We thank Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland for help with molecular modeling methods.

This work was partially supported by RFBR (grant No. 03-04-48623).

REFERENCES

Bonneau R., Tsai J. *et al.* (2001) Rosetta in CASP4: progress in ab initio protein structure prediction. *Proteins*, **Suppl 5**, 119–126.

Cho S.J., Lee M.G. *et al.* (2000) Crystal structure of *Escherichia coli* CyaY protein reveals a previously unidentified fold for the evolutionarily conserved frataxin family. *Proc. Natl. Acad. Sci. USA*, **97**, 8932–8937.

Edgar R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl. Acids Res.*, **32**, 1792–1797.

Granum P.E. (1997) *Bacillus cereus*. In Doyle M., Beuchat L., Montville T. (eds), *Fundamentals in Food Microbiology*. Washington, DC: ASM, pp. 327–336.

Kosinski J., Gajda M.J. *et al.* (2005) Frankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. *Proteins*, Suppl **7**, 106–113.

Kurowski M.A., Bujnicki J.M. (2003) GeneSilico protein structure prediction meta-server. *Nucl. Acids Res.*, **31**, 3305–3307.

Miles G., Bayley H. *et al.* (2002) Properties of *Bacillus cereus* hemolysin II: a heptameric transmembrane pore. *Protein Sci.*, **11**, 1813–1824.

Pierce B., Tong W. *et al.* (2005) M-ZDOCK: a grid-based approach for Cn symmetric multimer docking. *Bioinformatics*, **21**, 1472–1478.

Sali A., Blundell T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.

Song L., Hobaugh M.R. *et al.* (1996) Structure of staphylococcal alpha-hemolysin, a heptameric transmembrane pore. *Science*, **274**, 1859–1866.

Ye Y., Godzik A. (2004) FATCAT: a web server for flexible structure comparison and structure similarity searching. *Nucl. Acids Res.*, **32**, W582–585.

IDENTIFICATION AND STRUCTURE-FUNCTIONAL ANALYSIS OF THE SPECIFICITY DETERMINING RESIDUES OF THE ALPHA SUBUNITS OF THE PROTEOSOMAL COMPLEX

***Baryshev P.B.*[*1], *Afonnikov D.A.*[1, 2], *Nikolaev S.V.*[2]**
[1] Novosibirsk State University, Novosibirsk, 630090, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: pavel1983@ngs.ru

Key words: proteosome, protein-protein interactions, evolution, gene duplication, specificity determining positions

SUMMARY

Motivation: Proteosomes are polyenzymatic proteolytic structures that provide the degradation of the bulk of cytoplasmic proteins to oligopeptides. The proteosomal genes in the eukaryotes all arose by duplication of a single ancestral gene encoding the proteosomal subunits in the bacteria. The analysis of evolutionary events after duplication may be useful for discovering new information about proteosomal structural and functional properties.

Results: We confine our study here to the detection of the positions of the α-subunits whose amino acid substitutions are specific to particular subunits of the proteosomal alpha-rings. We detected a set of the α-subunit positions whose substitutions are specific to the genes that encode the various proteosomal subunits. It was demonstrated that these specific amino acid substitutions are the features of residues that form the subunit contacts in the α-ring of the proteosomes.

Availability: The proteosomal sequences, multiple sequence alignments and phylogenetic tree used in analysis are available upon request.

INTRODUCTION

According to the current concepts, the active moiety of the proteosome (20S) results in self-assembly of the subunits, a ring of seven α-subunits is assembled first, then a ring of β-subunits is built in (Kopp *et al*., 1997). It has been suggested that the order of the subunits in the proteosomal complex is fixed, i.e. each subunit in the ring occupies strictly defined place. Like in the case of self-assembly the proteosome, the order is defined by complementary interaction of the subunits, dependent on the spatial and physicochemical complementarity of the interacting parts of the macromolecules.

The evolutionary history of the α-subunit encoding genes in the eukaryotes is that they all arose by duplication of a single ancestral gene encoding the α-subunit in the bacteria (DeMartino, Slaughter, 1999); genome early during eukaryotic phylogeny. After duplication, as a result of a divergent evolution, each paralog gave rise to a group of orthologs, with each coding for 1 to 7 subunits that form the α-proteosomal ring in eukaryotes, including yeast and mammals (Bouzat *et al*., 2000). This model of evolution, based on the phylogenetic analysis of protein sequences, underlies the currently accepted

classification of the α-subunits in the paralogous groups. The model also implies that, after duplication, the α-subunit encoding genes kept accumulating mutations under selection pressure designed to maintain the stable ordering of this multi-subunit-structure (Nikolaev, Afonnikov, 2004). Thus, detection of such mutations and their analysis would give important information about how the features of this multisubunit structure might have formed.

We confine our study here to the detection of the positions of the α-subunits whose amino acid substitutions are specific to particular subunits of the proteosomal alpha-rings. To this end, we used the method implemented in the SDP program (Kalinina *et al*., 2004). We detected a set of the α-subunit positions whose substitutions are specific to the genes that encode the various proteosomal subunits. It was demonstrated that these specific amino acid substitutions are the features of residues that form the subunit contacts in the α-ring of the proteosomes.

MATERIALS AND METHODS

The sequences of the 20S proteosome subunits were retrieved from the SWISS-PROT database (Boeckmann *et al.*, 2003). An additional database search of homologous sequences was done using the BLASTP program (Altschul *et al.*, 1997). As a result, additional members of the proteosomal α-subunit family were chosen.

The CLUSTALW program (Thomson *et al*., 1994) was applied for the multiple alignment of the sequences. Analysis of the phylogenetic tree built by CLUSTALW program allowed us to assign the α-subunit sequences to paralogous groups. The groups include sequences from species exemplifying all the seven subunits types. After group assignment, the yielded multiple alignment was used to assess the conservation/variability at protein positions.

To define the positions with the subunit-specific mutations we used the SDPPred program (Kalinina *et al*., 2004). To estimate the significance of the positions the SDPPred uses the mutual information values. The values express the relation between the amino acid type at a given position and the index of paralog group (in our case it was the subunit index in the proteosomal ring from A to G) calculated as $I_i = \sum_{x,y} f(x_i, y) \cdot \log \frac{f(x_i, y)}{f(x_i) \cdot f(y)}$, where $f(x_i)$ is the occurrence rate of the amino acid x at the position i of the multiple sequence alignment, $f(y)$ is the fraction of the proteins assigned to the paralog group y, $f(x_i, y)$ is the occurrence frequency of the amino acid type x at the position i of the proteins in the paralog group y.

The identified positions were mapped to the proteosomal 3D structure (Unno *et al.*, 2002; PDB identifier 1IRU). The program iMoltalk (http://i.moltalk.org) was used to determine inter-subunit contact positions.

RESULTS AND DISCUSSION

In the course of the preliminary search, we choose 193 sequences of 35 species, of which 4 belonged to bacteria, 7 to archea and other to eukaryotes. Based on the surveyed phylogenetic tree we choose 7 paralogous groups. Each group corresponded to the homologs of the particular subunit of the proteosomal α-ring. This group assignment is based on the idea that the family genes resulted from single or series of duplications followed by sequence divergence. Therefore, homology among the sequence of different members of the family are supposed the result of sequence divergence after specific

events (the orthologous genes) or after the duplication of genes within an ancestral species (the paralogous genes) (Bouzat *et al*., 2000).

A SDPPred program was used to analyze the set of aligned sequences of the α-subunit sequences chosen as described above. SDPPred detected 25 positions at which the amino acid residues were conserved among the orthologs and different among the paralogous groups. Using the iMoltalk program we obtained that every subunit has at least 8 contact positions with the other α-subunits (positions 48, 53, 54, 72, 105, 209, 215, 224 of the multiple sequence alignment). Obtained contact positions are shown on Fig. 1. Moreover, certain amino acid residues form associations with the β-subunits. Conservatism of the remaining positions allow us to assume their importance for formation of the spatial proteosome structure. Mutations at such positions can result in incorrect folding of subunits and disrupt proteosomal complex formation. The chi-square test was applied to determine the significance of the positions detected by SDPPred program and the structural data (Table 1). We estimated the significance between the specific fixation the amino acid residues with respect to the subunit index, with the involvement of such position in the protein-protein interface for each of the α-subunit protein chain. The results shows, that the significance varies between subunits approaching the 90 % significance level.

The results suggested that during early phylogenesis, duplication in the subunit sequences was followed by mutations of residues that forms protein-protein interface and were important for the specific packing of subunits in the proteosomal machine.

Table 1. List of positions assigned as specificity-determining by SDPPred program and amino acids specific to the subunit sequence in the structure 1IRU. CP: 15 positions forming inter-subunit contacts; FP: 10 positions are not in contact with other subunits, likely responsible for the formation of the secondary and tertiary subunit structure

No.	SDP	Paralogous groups						
	CP	Chain A	Chain B	Chain C	Chain D	Chain E	Chain F	Chain G
1	48	S	E	S	S	R	R	G
2	53	R	F	S	R	R	N	L
3	54	H	S	R	A	G	D	S
4	55	I	L	T	I	V	V	A
5	61	E	S	E	D	E	Q	D
6	72	K	A	E	E	E	E	K
7	76	Q	G	H	K	L	Q	N
8	105	G	A	A	G	T	S	C
9	208	S	M	K	K	A	T	V
10	209	Q	Q	Q	Q	L	Q	H
11	215	A	G	G	N	−	Y	S
12	224	M	V	K	R	S	R	V
13	260	E	E	K	−	−	D	V
14	290	F	W	W	W	C	C	Y
15	301	K	N	N	S	A	R	A
	FP	Chain A	Chain B	Chain C	Chain D	Chain E	Chain F	Chain G
1	63	R	K	R	H	R	R	R
2	65	Y	V	Y	F	F	H	F
3	113	T	T	A	V	V	A	−
4	116	Q	E	R	K	R	R	E
5	159	M	Y	V	F	M	I	V
6	358	A	A	V	A	I	A	I
7	362	L	L	T	V	V	T	V
8	363	S	K	M	V	M	L	H
9	364	T	E	D	Q	E	P	D
10	410	K	G	V	L	N	E	N

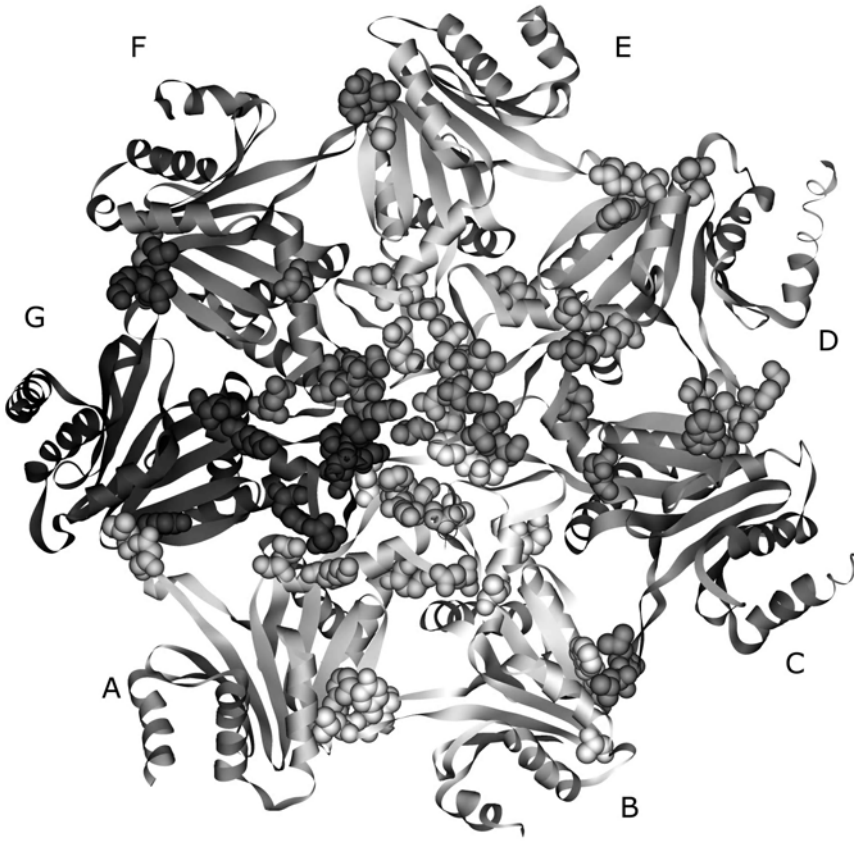


Figure 1. Spatial structure of the α-subunit ring. Subunit indices A-G are shown. SDP residues are shown in ball representation.

ACKNOWLEDGEMENTS

The work is supported by the U.S. Civilian Research & Development Foundation for the Independent States of the Former Soviet Union (CRDF) within the Basic Research and Higher Education Program (Y1-B-08-20, REC-008), the Ministry of Education grant РНП.2.1.1.4935, Russian Foundation for Basic Research (05-04-49141-a).

REFERENCES

Altschul S.F., Madden T.L., Schaffer 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.

Boeckmann B., Bairoch A., Apweiler R., Blatter M.-C., Estreicher A., Gasteiger E., Martin M.J., Michoud K., O'Donovan C., Phan I., Pilbout S., Schneider M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res.*, **31**, 365–370.

Bouzat J.L., McNeil L. K., Robertson H.M., Solter L.F., Nixon J.E., Beever J.E., Gaskins H.R., Olsen G., Subramaniam S., Sogin M.L., Lewin H.S. (2000) Phylogenomic analysis of the α proteosome gene family from early-diverging eukaryots. *J. Mol. Evol.*, **51**, 532–543.

DeMartino G.N., Slaughter C.A. (1999) The proteosome, a novel protease regulated by multiple mechanisms. *J. Biol. Chem.*, **274**, 22123–22126.

Kalinina O.V., Mironov A.A., Gelfand M.S., Rakhmaninova A.B. (2004) SDPpred: a tool for prediction of amino acid residues that determine differences in functional specificity of homologous proteins. *Nucl. Acids Res.*, **32**, W424–W428.

Kopp F., Hendil K.B., Dahlmann B., Kristensen P., Sobek A., Uerkvitz W. (1997) Subunit arrangement in the human 20S proteosome. *Proc. Natl. Acad. Sci. USA*, **94**, 2939–2944.

Nikolaev S.V., Afonnikov D.A. (2004) Inter-subunit contacts of the proteosomal alpha-subunits as determinants of paralog groups. *Proceedings of the Fourth International Conference on Bioinformatics of Genome Regulation and Structure*, Novosibirsk, Russia, 2004, **1**, p. 319–322.

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.

Unno M., Mizushima T., Morimoto Y., Tomisugi Y., Tanaka K., Yasuoka N., Tsukihara T. (2002) The structure of the mammalian 20S proteosome at 2.75 A resolution. *Structure*, **10**, 609–618.

CLUSTERING ANALYSIS OF CONFORMATIONAL STATES OF SHORT OLIGOPEPTIDES

Batsianovsky A.V.*, *, Vlasov P.K.
Engelhard Institute of Molecular Biology, RAS, Moscow, Russia; * Moscow State University, Moscow, Russia
* Corresponding author: e-mail: suner_s@mail.ru

Key words: peptide structure, conformation, clusters analysis

SUMMARY

Motivation: A goal of presented investigation is analysis of conformation combinations in proteins. There are no doubt about that conformations of amino acids are combining with some preferences (Vlasov *et al.*, 2005). At now there is sufficient material to analyze thoroughly structural state of tri- and tetrapeptides as a minimal symmetrical unit of regular structures of peptide backbone. It is tentative to reveal the pattern of the alternation of conformational states characterizing by different spiral symmetry.

Results: We have performed clustering analysis, using standard clustering method "K-means", to determine for most part of amino-acid residues the frequency of distribution of sequences of dihedral angles (φ- and ψ-angles), which are typical for protein helical conformations. It was found that the alternations amino-acid residues with different symmetry (3.6 in α-helix, 2 in β-strand) occur quite rarely in comparison to the alternations with similar symmetry. Influence of electrostatic in respect of limitations onto secondary structure alternation is discussed.

Availability: none

INTRODUCTION

Major part of a globule is formed by a set of symmetrical structures as α-helix, β-strand and others. It is obvious that formation a symmetrical network of intrinsic bonds including possible hydrogen bonds between atoms in main chain is of important for organization of protein structure and is favorable in respect of globule energetic. These symmetrical structures play essential role in protein folding. At the same time some asymmetrical elements are common in any protein structure. Fragments of such a kind wait their classification and functional annotation.

One can anticipate that majority of specific conformations must be combination of typical conformations belonging to allowed regions in Ramachandran map. These typical conformations correspond to secondary structures in general but other structural elements also possible.

Two questions arise in connection with sequence of conformations study:

- How many conformations can be constructed as a combination of typical regular protein conformations using corresponding dihedral angles? After this the symmetrical features of conformational state will be determined.
- What about of different conformation sequences occurrence?

MATERIALS AND METHODS

The method of clustering can readily applied for separation of different structural patterns. In this work a set of dihedral angles in oligopetides of fixed length was chosen as input data for clustering. Wide-spread method of K-Means implemented in mathematical package Statistica v 6.0 was used to solving the problem. Special program has been developed for interpretation results of clustering. This approach has been applied to the data on dihedral angles in protein structures from PDBSelect. This set of protein data is useful for many aims and may be treated as representative and sufficient narrow possessing fast computations. It should be noted that these data base includes nonhomological proteins (identity is not more 30 %). Method of ε-networks was applied for improving the quality of results. This method is convenient for elimination regions with rarely density of population.

RESULTS AND DISCUSSION

Results of clustering present division of set of structures into classes. For each class we estimated average values of operation parameters, deviation of these parameters inside the clusters, the number evinces in cluster. Unfortunately, we could not achieve the absolute division. It is likely in connection with the character of data and space operation factors, and clustering method peculiarities. However, we could distinguish conservative clustering in different divisions. These clusters conserve their characteristics in all divisions. This evidently implies that these clusters are correct and reflect the real structures in proteins.

Conformations of oligopeptides from conserved clusters are combinations of conformational states of typical conformations as expected. Structures with non-typical conformational states do not form constant clusters. For example, we have found some clusters with β-turns, but all of them were unstable. As a result it is possible to descript these stable structures in terms of typical conformations denoting the conformation state of amino acid by region of typical conformation.

Frequencies of these structures show strong nonregularity. This effect can be explained on the basis of symmetry mechanism of compensation of electrostatic interactions. For example, pairs of adjacent amino acid residues in β-conformation raise frequencies of occurrence of the structure (symmetry in β-strand is 2), at the same time very small fragments (included one or residues) in α-conformation especially in the middle of oligopeptide reduce the frequency (symmetry in α-helix – 3.6). There is prohibition of some structural types. Such structures as **β-α-β-α**, **α-β-α-β**, **β-α-α-β** are absent in all divisions. There is no compensator electrostatics in these structures. At last, structures **α-α-β** and **α-α-β-α** occur only in some divisions. We obtained the list of stable clusters of tripeptides and tetrapeptides:

Table 1. List of stable clusters of tripeptide and tetrapeptide structures. The symbol β signify conformational state corresponding to region of β-strand in the Ramachandran map, α – α-helix, lα – left α-helix. The total number of structures in this sample is 637 thousands. Average deviation from indicated numbers in cluster is lower than 3 thousands for the large clusters, and 2 thousands for small ones

Structure	No, thous	Structure	No, thous
β-β-β	~ 130	β-β-β-β	~93
α-α-α	~220	α-α-α-α	~190
α-β-α	~17	α-β-β-β	~23
α-β-β	~32	β-α-α-α	~23
β-α-β	~21	β-β-β-α	~24
β-α-α	~35	β-α-β-β	~20
β-β-α	~32	β-β-α-α	~18

Structure	No, thous	Structure	No, thous
β-β-lа	~14	β-β-α-β	~20
lа-β-β	~17	α-α-β-β	~15
β-lα-β	~8	α-α-α-β	~19
α-α-lα	~13	α-β-α-α	~10
		α-β-β-α	~9
		β-β-β-lα	~11
		β-lα-β-β	~15
		lα-β-β-α	~5

A statistical approaches as operation of clustering is useful for finding of general rules that control structure formation in short oligopeptides. Clustering operation is partly complicated by noise effect through all Ramachandran plot, and application of ε-networks cannot overcome the noise effect. Maybe in turn the space of operation parameters is not natural for determination of structural types. However, in spite to this negative factors it was achieved clustering in reasonable and prominent groups of structures. The results demonstrate that there are no equal frequencies of various combinations of elementary conformation states. On of possible explanation consists in taking into account the electrical dipole moment of amino acid residues and possible electrostatic compensatory effects in this aspect.

REFERENCES

Vlasov P.K., Vlasova A.V., Tumanyan V.G., Esipova N.G. A tetrapeptide-based method for proline II-type secondary structure prediction. *Proteins*, **61**, 763–768.

"STRANGE KINETICS" OF UBIQUITIN FOLDING: INTERPRETATION IN TERMS OF A SIMPLE KINETIC MODEL

Chekmarev S.F.*[*1], *Krivov S.V.*[2], *Karplus M.*[2, 3]
[1] Institute of Thermophysics, SB RAS, Novosibirsk, 630090, Russia; [2] Laboratoire de Chimie Biophysique, ISIS, Université Louis Pasteur, 67000, Strasbourg, France; [3] Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA, e-mail: marci@tammy.harvard.edu
[*] Corresponding authors: e-mail: chekmarev@itp.nsc.ru

Key words: ubiquitin, folding, folding times, rate constants, kinetic model, experiment

SUMMARY

Motivation: Interpretation of protein folding experiments in terms of the transitions between characteristic states of the system may allow valuable insight into the folding mechanism.

Results: The ubiquitin mutant Ub*G folding experiments of Sabelko *et al*. (1999), in which "strange kinetics" were observed, are interpreted in terms of a simple kinetic model. A minimal set of states consisting of a semi-compact globule, two off-pathway traps, and the native state are included. Both the low and high temperature experiments of Sabelko et al. are fitted by a system of kinetic equations determining the transitions between these states. It is possible that cold and heat denaturated states of Ub*G are the basis of the off-pathway traps. The fits of the kinetic model to the experimental results provides an estimate of the rate constants for the various reaction channels and show how their contributions vary with temperature. Introduction of an on-pathway intermediate instead of one of the off-pathway traps does not lead to agreement with the experiments.

INTRODUCTION

Recent progress in experimental studies of protein folding on millisecond and faster time scales has increased the interest in the existence and role of intermediates in the folding process (Ferguson, Fersht, 2003). Ubiquitin is a benchmark system for protein folding studies due to its structural and physical properties (Went *et al*., 2004). To provide a fluorescence probe for monitoring folding of ubiquitin, Khorasanizadeh *et al*. (1993) replaced the largely buried Phe45 by a Trp. In the recent temperature-jump experiments of Sabelko *et al*. (1999), which we analyze here, a double mutant (F45W, V26G) of human ubiquitin (Ub*G) was used; the V26G mutation was added to destabilize core contacts and a critical helix to make cold denaturation possible. Sabelko *et al*. found that the time-dependent population of unfolded states varied with temperature from a double-exponential distribution at T = 2 °C to a stretched-exponential one at T = 8 °C. The interest of the latter is that it occurs on warming rather than cooling, and the high temperature spectrum of characteristic times is (quasi-) continuous and spans a broad region, corresponding to what has been referred to as "strange kinetics" (Shlesinger *et al*., 1993). In protein folding such kinetics have been associated with downhill folding through an array of intermediates separated by relatively low free energy barriers (e.g., Skorobogatiy *et al*., 1998). Sabelko *et al*. (1999) adduced arguments in favor of a

downhill folding scenario at T = 8 °C but also indicated that the kinetics at this temperature is well fitted by a three-exponential distribution. Thus, the presence of long-living intermediates could not be ruled out. Stretched-exponential distributions are convenient for approximation of experimental data, because they require just two parameters, a time constant and a power of the stretching exponent. A shortcoming of this approach is that the relation between the energy landscape and the parameters is not direct; i.e. knowledge of the parameters gives little information about specific features of the landscape. Multi-exponential distributions require more fitting parameters. However, if these distributions are related to a specific model and represent solutions of a system of kinetic equations, valuable insight into the folding process can be obtained. Recently, we simulated the Monte Carlo folding kinetics of a 27-bead square lattice protein model and showed that the results could be described by a simple kinetic model with off-pathway intermediates (Chekmarev *et al*., 2005). In this work we use a similar model to interpret the data on Ub*G folding.

KINETIC MODEL

We consider three kinetic schemes of the folding process, which include a semi-compact globule, off- and on-pathway intermediates and the native state. The schemes differ in the number and type of the intermediates involved: scheme #1 includes one off-pathway trap, scheme #2 two separate off-pathway traps, and scheme #3 one off-pathway trap and one on-pathway intermediate, which connects the globule and the native state. Given the rate constants, the time-dependent populations of the states are determined by the systems of linear kinetic equations. Following the experimental conditions of Sabelko *et al*. (1999), we do not include the fully extended (denaturated) state of the system in consideration. Also, since the experimental population of unfolded states does not show a tendency to stabilization at long times, we infer that native state unfolding is negligible. Therefore, the folding time is equated to the first passage time.

Solving the systems of kinetics equations yields the time-depending populations of the states as functions of the rate constants, among them, the population of the unfolded states, which was measured by Sabelko *et al*. (1999). Then the fit of the theoretical solution to experimental distribution can be used to estimate the rate constants.

RESULTS AND DISCUSSION

Testing the previously mentioned kinetic schemes against the experimental results of Sabelko *et al*. (1999), we have found that both the low and high temperature experiments (i.e. at T = 2 °C and T = 8 °C) are well described within the framework of scheme #2. However, at low temperature (T = 2 °C) only one trap is essential and scheme #2 reduces to scheme #1. Table 1 shows the corresponded values of the fitted rate constants, and Fig. 1 and 2 compare the theoretical solutions with the experimental results. Introduction of an on-pathway intermediate instead of one of the off-pathway traps (scheme #3) does not lead to agreement with the experiments.

Table 1. Ub*G folding: Rate constants $r_{\beta\alpha}$ (μs^{-1}) and waiting times $\tau_{\beta\alpha} = 1/ r_{\beta\alpha}$ (μs, in brackets) for the transitions from state α to β. Subscripts g, d1, d2 and f stand for the semi-compact globule, off-pathway traps 1 and 2, and the native state, respectively

	$r_{d1,g}$ ($\tau_{d1,g}$)	$r_{g,d1}$ ($\tau_{g,d1}$)	$r_{d2,g}$ ($\tau_{d2,g}$)	$r_{g,d2}$ ($\tau_{g,d2}$)	$r_{f,g}$ ($\tau_{f,g}$)
T = 2°C	8.0×10^{-3} (125)	2.4×10^{-4} (4105)	4.1×10^{-9} (2.4×10^{9})	1.7×10^{-2} (60)	3.7×10^{-2} (27)
T = 8°C	2.0×10^{-3} (510)	5.6×10^{-4} (1799)	2.0×10^{-2} (51)	8.0×10^{-3} (125)	1.9×10^{-2} (52)

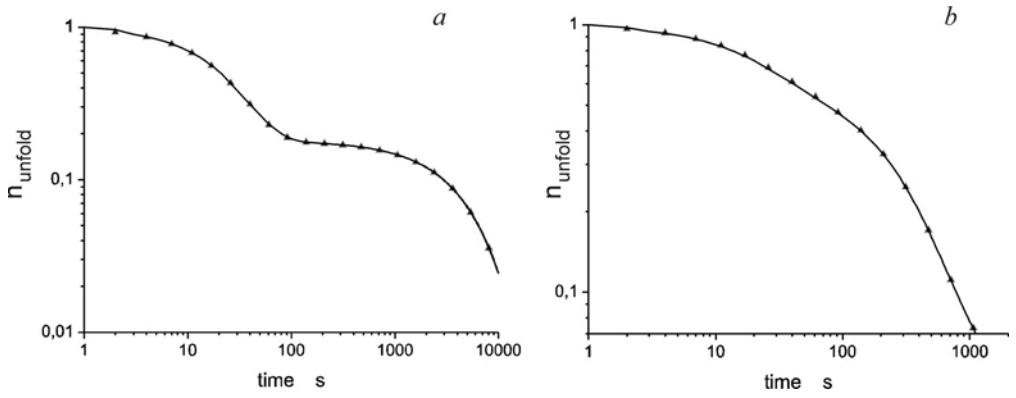


Figure 1. Populations of the unfolded states of Ub*G at T = 2 °C (panel *a*) and T = 8 °C (panel *b*). The triangles correspond to the experimental data of Sabelko *et al*. (1999), and the solid line to the theoretical solution with the rate constants from Table 1.

With a knowledge of the forward and backward rate constants (Table 1), it is possible to calculate the free energies of the off-pathway traps with respect to the globule. Assuming detailed balance, $r_{j,\,i}/r_{i,\,j} = \exp\left[-\left(F_j - F_i\right)/k_B T\right]$, where F_α is the free energy of state α, and k_B is the Boltzmann constant, we found $F_{d1} - F_g \approx -1.9$ and $F_{d2} - F_g \approx 9.6$ at T = 2 °C, and $F_{d1} - F_g \approx -0.7$ and $F_{d2} - F_g \approx -0.5$ at T = 8 °C, with the energy measured in kcal/mol. An additional assumption about the value of the prefactor $A_{j,\,i}$ in the Arrhenius equation, $r_{j,\,i} = A_{j,\,i}\ \exp\left(-\Delta F_{j,i}/k_B T\right)$, made it possible to estimate tentative values of the effective free energy barriers between the states, $\Delta F_{j,\,i}$. For this, we assume that the $A_{j,\,i}$ are equal to the frequently used :speed limit: of $1\mu s^{-1}$, characteristic of small proteins (Kubelka *et al*., 2004). The results are shown in Table 2, and they are similar in magnitude to the free energy barriers obtained for a set of small proteins (Kubelka *et al*., 2004). Of course, the actual value of the barrier depends on the choice of the pre-exponential factor. The observed large variation of the effective free energies with temperature is indicative of the complexity of the system and the temperature dependence of the reduced energy landscape.

In accord with the variation of the free energy surface with temperature, which was proposed for ubiquitin (Sabelko *et al*., 1999), it is possible that the cold and heat denaturated states of Ub*G are involved in the off-pathway traps. If so, when one passes from T = 2 °C, where cold denaturation dominates, to T = 8 °C, which corresponds to the optimum native stability (Sabelko *et al*., 1999), the role of the cold denaturated state in the folding process would decrease and that of the heat denaturated state would increase. Alternatively, a trap could be associated with a state having excess helix relative to the native state (M. Gruebele, personal communication). We hope the present analysis will stimulate further studies of ubiquitin, a model system for which a definitive folding mechanism is likely to be determined by additional experimental and theoretical analyses.

Table 2. Ub*G folding: Effective free energy barriers between the characteristic states (kcal/mol). Subscripts g, d1, d2 and f are as in Table 1

	ΔF_{d1g}	ΔF_{gd1}	ΔF_{d2g}	ΔF_{gd2}	ΔF_{fg}
T = 2 °C	2.6	4.6	11.8	2.2	1.8
T = 8 °C	3.5	4.2	2.2	2.7	2.2

ACKNOWLEDGEMENTS

We thank M. Gruebele for helpful comments and P. Maragakis for useful discussion. This work was supported in part by a grant from the CRDF (RUP2-2629-NO-04). The research at Harvard was supported in part by a grant from the National Institutes of Health. S.F. Ch. also acknowledges support from the Russian Foundation for Basic Research (# 06-04-48587). The material for this report is taken mainly from the article: S.F. Chekmarev, S.V. Krivov and M. Karplus, J. Phys. Chem. B, v.110 (in press).

REFERENCES

Chekmarev S.F., Krivov S.V., Karplus M. (2005) Folding Time Distributions as an Approach to Protein Folding Kinetics. *J. Phys. Chem. B*, **109**, 5312–5330.

Ferguson N., Fersht A. R. (2003) Early events in protein folding. *Curr. Opin. Struct. Biol.*, **13**, 75–81.

Khorasanizadeh S., Peters I.D., Butt T.R., Roder H. (1993) Stability and folding of a tryptophan-containing mutant of ubiquitin. *Biochemistry*, **32**, 7054–7063.

Kubelka J., Hofrichter J., Eaton W.A. (2004) The protein folding 'speed limit'. *Curr. Opin. Struct. Biol.*, **14**, 76–88.

Sabelko J., Ervin J., Gruebele M. (1999) Observation of strange kinetics in protein folding. *Proc. Natl. Acad. Sci. USA*, **96**, 6031–6036.

Shlesinger M.F., Zaslavsky G.M., Klafter J. (1993) Strange Kinetics. *Nature*, **363**, 31–37.

Skorobogatiy M., Guo H., Zuckermann M. (1998) Non-Arrhenius modes in the relaxation of model proteins. *J. Chem. Phys.*, **109**, 2528–2535.

Went H.M., Benitez-Cardoza C.G., Jackson, S.E. (2004) Is an intermediate state populated on the folding pathway of ubiquitin? *FEBS Letters*, **567**, 333–338.

A METHOD TO ASSESS CORRECT/MISFOLDED STRUCTURES OF TRANSMEMBRANE DOMAINS OF MEMBRANE PROTEINS

***Chugunov A.O.*[*1, 2], *Novoseletsky V.N.*[1, 3], *Efremov R.G.*[1]**
[1] Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, RAS, GSP Moscow, 117997, Russia;
[2] Department of Bioengineering, Biological Faculty, M.V. Lomonosov Moscow State University, 119899, Moscow; [3] Moscow Institute of Physics and Technology (State University), 141700, Russia
[*] Corresponding author: e-mail: volster@nmr.ru

Key words: membrane proteins, transmembrane domain, scoring function, residue environment, G-protein coupled receptors, rhodopsin

SUMMARY

Motivation: Integral membrane proteins (MP) are pharmaceutical targets of exceptional importance since more than 50 % of currently marketed drugs target these objects. Due to technical difficulties, modern experimental methods often fail to determine 3D structure of MPs. Computational methods for modeling MPs structure and assessment of these models' quality may be very helpful in this case.

Results: We propose a novel method for quantitative estimation of the transmembrane (TM) domains models' quality. The approach is based on the concept of environmental profile. A non-redundant set of 26 high-resolution X-ray structures of α-helical TM domains is used to define five classes of residues' environment, considering polarity of nearest protein surrounding and accessibility for a given residue. Residues' preferences for each environment class are calculated. The main results are: (1) The proteins length correlates with the proposed scoring function values, defining a way to differentiate "well-folded" structures from misfolded ones; (2) The method efficiently delineates crystallographic structure of visual rhodopsin both in a set of twelve its computer models, containing certain errors and ensemble of artificially generated misfolded structures of rhodopsin; (3) Photosynthetic MPs demonstrate different score-length dependency, suggesting distinct packing characteristics for these proteins.

INTRODUCTION

Integral membrane proteins (MP) are objects of special biological and pharmaceutical importance, establishing every cell's communication with the rest of the world, including signal transduction, light absorption and formation of TM potential. A very large and important class of MPs, G-protein coupled receptors (GPCRs), is a target for > 50 % of currently marketed drugs. Unfortunately, possibilities of modern experimental techniques for MPs 3D structure determination are far under pharmaceutical industry (e.g., structure-based drug design) requirements. Current proportion of MPs in Protein Data Bank (PDB) is less than 1 %, whereas every sequenced to date genome encodes 15–30 % of membrane proteins. To overcome this discrepancy, development of computational methods for modeling MPs structure and assessment of these models' quality is believed to be very helpful.

Many efforts have been made to understand the principles of structural organization of MPs, but the problem is yet to be solved. What are the differences in their packing and structure as compared to soluble proteins? Some methods designed for MPs structure prediction utilize sequence statistics or more general characteristics (e.g. protein packing density), but only very few of them use high-resolution structural data on residues' environments.

METHODS AND ALGORITHMS

Creation of membrane proteins database. We used MPs structures with primarily α-helical TM domains, determined by high-resolution (< 3.5 Å) X-ray crystallography. The training set contains 26 structures of proteins that have no sequence homology to each other. A separate set contains 6 structures of photosynthetic proteins. All protein structures were aligned along the membrane normal (hereinafter, *Z* axis) in order to place cytoplasmic sides of plasma MPs and matrix sides of inner mitochondria MPs to $Z < 0$ area, and *vice versa*. Optimal *Z* position of the entire structure and the thickness of "TM" hydrophobic layer were determined by finding solvation energy minimum using implicit membrane-water environment model (Efremov *et al.*, 2000). Only α-helical residues (as determined by DSSP (Kabsch, Sander, 1983)) within "optimal" hydrophobic layer (27 Å in average) plus 5 Å at each side were selected for the study.

"Membrane Score" calculations. In order to characterize the environment of a particular residue, we used fractions of full residue's surface that face polar and non-polar atoms of *other* TM α-helices, respectively:

$$Fp^1 = Fp' - Fp^0 \equiv \frac{Sp}{S^0} - \frac{Sp^0}{S^0}, Fnp^1 = Fnp' - Fnp^0 \equiv \frac{Snp}{S^0} - \frac{Snp^0}{S^0}, \quad (1)$$

where Fp' and Fnp' are areas of polar (Sp) or non-polar (Snp) contacts, divided by the residue "self" area in Gly-Res-Gly motif. Fp^0 and Fnp^0 are the corresponding values for *isolated* TM helices. The difference between them is a measure of interhelical effects. Given the $Fp^1 \times Fnp^1$ distributions for *i* residue (Fig. 1*a*) and membrane environments scheme with parameters *a*, *b* and tg*α* (Fig. 1*b*), we define a membrane scoring function:

$$MemScore_{ij} = \ln(\frac{P_{ij}}{P_j}), \quad (2)$$

$$TotalMemScore = \sum_{ij} N_{ij} \ln(\frac{P_{ij}}{P_j}), \quad (3)$$

where P_{ij} is a probability to find residue *i* in environment *j*, and P_j is a probability to find *any* residue in environment *j* (*j* occupancy). The total score was calculated as in (3), where N_{ij} is a number of residues *i* in class *j*. Corresponding value in (2) was never zero and was set to unity if there were no *i* residues in *j* class, but if it was zero in (3), the entire term was not considered. Exact *a*, *b* and tg*α* values were determined in order to maximize the total score value (3) for the whole training set.

Rotameric test. We generated an ensemble of conformations of visual rhodopsin, where every TM α-helix was rotated around its axis with increment 90°, resulting in 16384 (4^7) rotameric conformations. A simple energy minimization was applied to avoid sterical clashes.

RESULTS AND DISCUSSION

To establish a method for assessment of MPs packing quality, we introduced two residues' environment characteristics, namely fractions of full residue surface that are in contact with polar and non-polar atoms of *other* TM α-helices, Fp^1 and Fnp^1, respectively (see Methods for details). For the whole training database (see Methods) we obtained distributions of these parameters for each residue type, as shown in Fig. 1*a*.

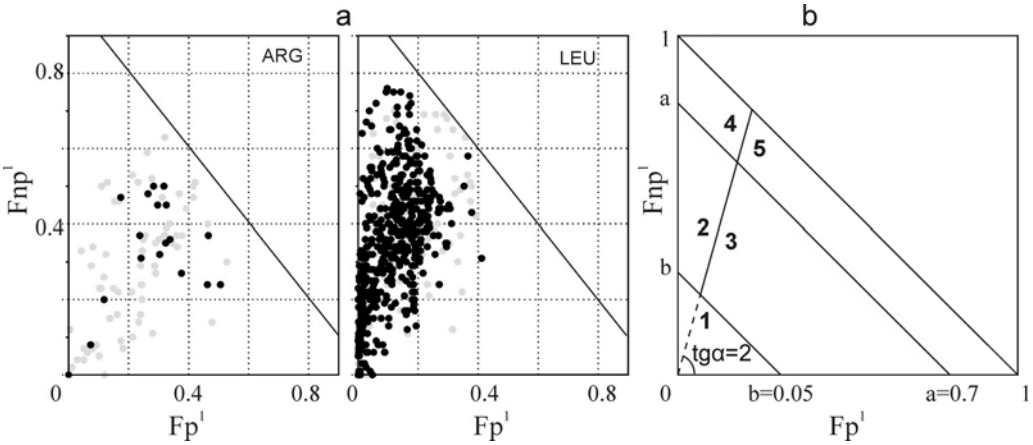


Figure 1. Residues' environmental characteristics distributions derived from "training" database for Arginine and Leucine (*a*). Black circles correspond to the "central" residue location ($|Z| < 15$ Å), and gray – to the interfacial. Proposed scheme for membrane environment classes (*b*). Class 1 corresponds to exposed one, classes 2 and 3 – to intermediately buried, 4 and 5 – to buried. Classes 2 and 4 correspond to non-polar environment, whereas 3 and 5 – to polar one.

As one can see, there are much more leucine residues in TM domains than arginines. This demonstrates strong preference of these residues to "central" and to interfacial locations, respectively. Also, most of leucines are situated in non-polar environment (close to "non-polar" Fnp^1 axis) and arginines – in polar. In $Fp^1 \times Fnp^1$ coordinates the proximity to zero means high accessibility for membrane milieu (high values of accessible solvent area, ASA), whereas location near the $Fp^1 = 1 - Fnp^1$ line means considerable burial ($ASA \approx 0$). Based on these observations, we propose the following scheme for definition of environmental classes for TM α-helical domains of proteins (Fig. 1*b*). The scores were calculated for each combination of residue type and membrane class (not shown), enabling assessment of the quality of the whole structures.

In Fig. 2 the membrane scores for proteins from the training set are plotted against TM domain length. It is seen that, there is a good correlation between them. This enables differentiation between correct (e.g. crystal) and misfolded structures. To test the possibility, we chose from public domain 12 computer models of bovine visual rhodopsin, built prior to release of its crystal structure (Palczweski *et al.*, 2000), and compared them in terms of membrane score values. As seen in Fig. 2, all of them lie below the crystal structure, and those that have been built in a fully automatic manner (e.g., at Swiss-Prot, MODBASE, GPCRDB servers), score much lower than the carefully optimized ones (data not shown). Also, a notable correlation exists between model's deviation from the crystal structure (in terms of r.m.s.d.) and score impairment (not shown). It was noticed, that photosynthetic proteins demonstrate different score-length dependency, suggesting distinct packing characteristics for them.

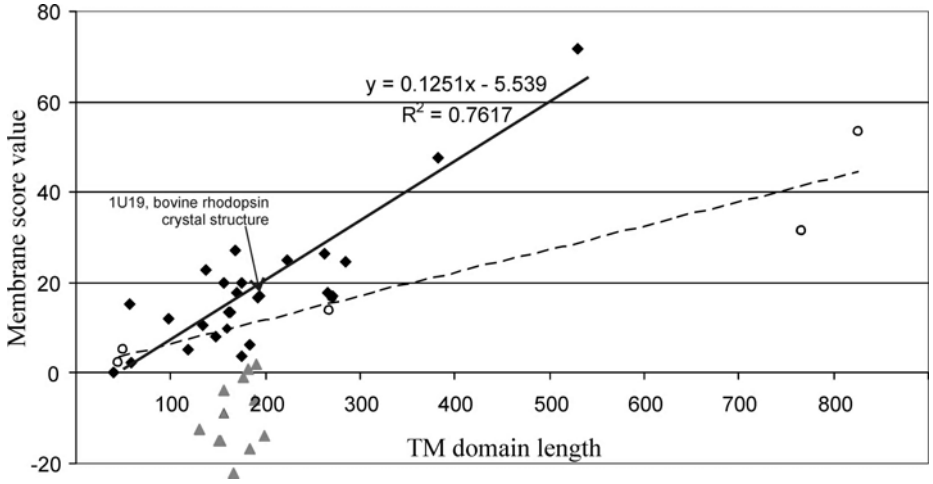


Figure 2. Membrane score values as function of TM domain length for training set (black diamonds), photosynthetic proteins (white circles) and computer models of visual rhodopsin (gray triangles).

In order to further validate our method's possibility to distinguish correct and misfolded structures, we generated more than 16000 rhodopsin's rotameric conformations (see Methods), vast majority of which are believed to be misfolded. For this ensemble, we compared the ability of our method to rank the crystal structure among it, with the results obtained using the well-known Eisenberg's method (Bowie *et al.*, 1991), conceptually close to ours, but parameterized for globular proteins. As seen from Fig. 3, "classical" method, very good for soluble proteins, is unable to mark out the crystal structure, whereas the present membrane-tuned method performs rather well.

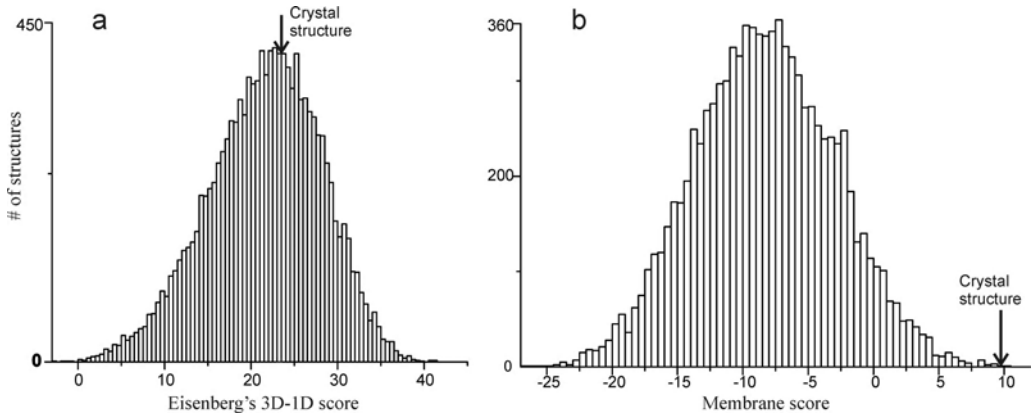


Figure 3. Distributions of values of scoring functions for ensemble of more than 16000 misfolded "rotameric" conformations of visual rhodopsin. Crystal structure position is shown with an arrow. *a* – eisenberg's 3D-1D scoring function; *b* – proposed "Membrane score" function.

To conclude, we have developed a novel method to estimate the packing quality of TM α-helical domains in proteins. We suppose that this method will be especially useful for GPCRs' models construction and optimization.

ACKNOWLEDGEMENTS

This work was supported by the Programme RAS MCB, by the Russian Foundation for Basic Research (grants 02-04-48882-a, 05-04-49346-a, 06-04-49194-a), by the Russian Federation Federal Agency for Science and Innovations (The State contract 02.467.11.3003 of 20.04.2005, grant SS-4728.2006.4).

REFERENCES

Bowie J.U., Luthy R., Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **5016**, 164–70.

Efremov R.G., Nolde D.E., Volynsky P.E., Arseniev A.S. (2000) Modeling of peptides in implicit membrane-mimetic media. *Mol. Simulation*, **24**, 275–291.

Kabsch W., Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **12**, 2577–637.

Palczewski K. *et al.* (2000) Crystal structure of rhodopsin: a G protein-coupled receptor. *Science,* **289**, 739–745.

DIRECT INFLUENCE OF UBIQUITYLATION ON A TARGET PROTEIN ACTIVITY: "LOSS-OF-FUNCTION" MECHANISM REVEALED BY COMPUTATIONAL ANALYSIS

Chernorudskiy A.L., Shorina A.S., Garcia A., Gainullin M.R.*
Nizhny Novgorod State Medical Academy, Nizhny Novgorod, Russia
*Corresponding author: e-mail: biochem@n-nov.mednet.com

Key words: ubiquitin, active site, domain, steric effects, 3D structure, computer analysis

SUMMARY

Motivation: Large volume of proteomic data concerning different protein ubiquitylation has been acquired in recent years. Particular ubiquitylation sites for these proteins were also identified. This allows us to analyze co-localization of ubiquitylation spots and functionally important protein domains, following the idea that ubiquitylation can directly affect functional activity of the proteins.

Results: Our results suggest that ubiquitylation can regulate functional activity of concerned proteins through direct steric effects.

INTRODUCTION

Ubiquitylation is a process of great importance for many vital cell functions. A covalent attachment of highly conserved small protein ubiquitin to another target protein significantly changes fate and functional state of target protein (Haglund, Dikic, 2005). The aim of the present work was a prediction of possible ubiquitylation effects on protein structural and functional properties through evaluation of steric effects of ubiquitin attachment. For this purpose we studied co-localization of proteins functionally important regions with all possible ubiquitin-modified lysine residues of following proteins: orothidine 5′-phospate decarboxylase (ODC), peroxisomal citrate synthase (PCS), X-linked inhibitor of apoptosis (XIAP) and vertebrate calmodulin (CaM). We take advantages of the computational approach to obtain 3D structures of these proteins and to analyze its rearrangement with ubiquitin molecule.

METHODS AND ALGORITHMS

Active site sequence revealing was performed using ScanProsite tool (http://www.expasy.org/tools/scanprosite/). Multiple sequence alignment was carried out using ClustalW. BioEdit Sequence Alignment Editor 7.0.5.2 was used for pairwise sequence alignment. Modelling of PCS 3D structure was conducted using Swiss-Model service (Automated Comparative Protein Modelling Server, http://swissmodel. expasy.org) and respective theoretical model was deposited in SwissModel Repository. Pairwise structure alignment was calculated interactively using the Combinatorial Extension method (http://cl.sdsc.edu/ce.html). Determination of amino acids participating

in intersubunit contacts was carried out using Protein-Protein Interaction Server v.1.5 (http://www.biochem.ucl.ac.uk/bsm/PP/server/). Swiss-PdbViewer program was used for 3D visualization and model analysis. Solvent accessible surface was calculated for a probe sphere of radius 1.4 angstroms.

IMPLEMENTATION AND RESULTS

Ubiquitylation of orotidine 5′-phosphate decarboxylase. Orotidine 5'-phosphate decarboxylase (EC 4.1.1.23) is responsible for conversion of orotidine 5′-monophosphate to uridine 5′-monophosphate, the last step in the *de novo* pyrimidine biosynthetic pathway. According to Peng and coworkers (Peng *et al*., 2003), the enzyme has 3 ubiquitin acceptor sites (Lys93, Lys209 and Lys253). We propose that ubiquitylation of the first one, Lys93, may lead to several consequences affecting enzyme activity. Lys93 participates in substrate binding during catalytic act, so its ubiquitylation will abolish substrate conversion. Moreover, ODC is active in a homodimeric form. Its both subunits participate in active site formation, and Lys93 plays a crucial role in this process. When introduced in the area of intersubunit contacts, ubiquitin will completely break these interactions and impede the dimerization. Both effects of Lys93 modification by ubiquitin mentioned will result in the complete loss of catalytic activity of the enzyme. This correlates with data of Smiley and Jones about loss of enzyme activity when Lys93 is substituted to cysteine (Smiley, Jones, 1992). We have analyzed the dimer surface and showed that Lys93 is buried in the intersubunit contacts area. As far as spatial accessibility of a lysine residue is a prerequisite for its ubiquitylation, we can postulate that Lys93 can be modified by ubiquitin only in the inactive monomeric form of ODC.

Another possible ubiquitylation site in ODC is Lys209. This residue is located on the protein surface nearby the “entrance” to active site cavity and can be ubiquitylated both in monomeric and dimeric forms of ODC. This modification may cause steric limitations for both active site accessibility for the substrate and for ODC dimerization ability, and therefore may also result in the decrease of enzymatic activity.

Lys253 is located rather far from functionally important areas of ODC molecule, so this residue ubiquitylation may hardly induce steric limitations.

Ubiquitylation of peroxisomal citrate synthase. Citrate synthase (EC 2.3.3.1) catalyzes the synthesis of citrate from oxaloacetate and acetyl-CoA. Ubiquitylation sites for PCS are Lys385 and Lys354 (Peng *et al*., 2003). There are no direct experimental data about spatial organization and 3D active site structure of PCS. High rate of conservativity of this enzyme in eukaryotes allowed us to carry out comparative modelling of PCS and to obtain reliable 3D structure. We have also identified amino acid residues forming PCS active site. A calculation of the solvent accessible surface for PCS 3D model reveals that active site residues form a pocket in protein globule surface. Analysis of spatial arrangement of acceptor lysine residues in PCS 3D structure revealed that both lysines are exposed on the surface of molecule, but located differently relative to active site. Conserved Lys385 residue is positioned on the globule surface very close to enzyme catalytic pocket, while Lys354 is located far from the active site. Such localization correlates with differences in rate of conservativity of two lysine residues.

Due to the fact of close proximity of Lys385 to the active site pocket and the flexibility of the ubiquitin C-terminal extension, we propose that ubiquitin attachment in this position may affect catalytic activity of the enzyme by reducing accessibility of PCS active site for substrates. Thus, ubiquitylation of PCS Lys385 leads to the same steric effects as observed in the case of ubiquitin attachment to Lys209 of ODC. On the contrary, we consider functional disturbances due to PCS Lys354 modification unlikely.

Similarly to ODC, citrate synthase is catalytically active in homodimer form. We consider dimeric mitochondrial citrate synthase (MCS) from a pig heart to be applicable

for analysis of interconnection between ubiquitylation and protein oligomerization, reasoning from high rate of similarity of two citrate synthases. We have shown that MCS lysine, homologous to Lys385 of PCS, is overlapped by the second subunit. Thus, two consequences can be derived from this observation. Similarly to Lys93 of ODC, ubiquitin attachment to PCS Lys385 is possible only in a monomeric form. Being ubiquitylated, single PCS subunit escapes from oligomerization.

Ubiquitylation of XIAP. XIAP is a powerful inhibitor of apoptosis, blocking both mitochondrial and Fas-mediated apoptosis pathways through direct binding and inhibition of various caspases. XIAP contains 3 BIR domains, which bind caspases, and a RING domain with ubiquitin ligase activity, responsible for protein autoubiquitylation. At present 2 ubiquitylation sites are identified for XIAP – Lys322 and Lys328, both localized within BIR3 domain of the protein (Shin *et al.*, 2003). BIR3 is responsible for interaction with initiator caspase 9 and Smac/DIABLO (mitochondria-derived activator of caspases). Analysis of interactions with XIAP revealed that aminoacids profiles of contact area are the same. Lys322 participates in contact area formation with caspase 9, but not with Smac/DIABLO. On the other hand, Lys328 participates in interactions with both XIAP partners. Therefore, ubiquitin binding to Lys322, which is major *in vivo* ubiquitylation site of XIAP, may block docking of caspase 9 to its BIR3 domain. It may also influence Smac binding, because Lys322 is situated very close to the contact area. The distance between this lysine and the nearest amino acid of docked Smac (Ala1) was calculated as 7,5 angstroms. As far as ubiquitin is a relatively large modifier and has a size of 44 angstroms, its introduction in this place may sterically disturb interaction with Smac.

Although Lys328 is a minor *in vivo* ubiquitin acceptor site, its modification may also block docking of both interaction partners on to XIAP.

Ubiquitylation of calmodulin. Calmodulin mediates the control of a large number of enzymes by Ca^{2+}. Its structure comprises 4 so-called EF hands, each of them can bind one Ca^{2+} ion. Among the enzymes to be stimulated by the calmodulin-Ca^{2+} complex are a number of protein kinases and phosphatases.

Calmodulin has 3 lysines that may be modified with ubiquitin. Probability of particular lysine ubiquitylation decreases in a row Lys13 – Lys21 – Lys94. It was experimentally shown that monoubiquitylation strongly decreases the biological activity of calmodulin by reducing its ability to activate phosphorylase kinase (Laub *et al.*, 1998). We have analyzed what mechanism can work in this case, using different CaM 3D structures from PDB. These structural data clearly show that CaM undergoes considerable conformational changes during interaction with its binding partners. We propose that ubiquitin attachment can limit conformational flexibility of CaM and therefore impair its ability to form complexes with interaction partners.

DISCUSSION

Computational analysis performed shows several direct effects of ubiquitylation on the function of target protein. Firstly, ubiquitylation can sterically block protein functional domains/active site or cause accessibility limitations, as for the cases of ODC and PCS. Secondary, it causes steric disturbances for homo-oligomerization, as it was shown for ODC and PCS. Then, it also influences heterologic protein interactions, impeding concurrent binding of target protein with its partners, as for XIAP and CaM. Furthermore, interaction with partner proteins can be disturbed due to limitations of conformational flexibility, as it was observed for CaM. Any of these effects will result in a decrease of target protein activity. Thus, we suggest a new “loss-of-function” mechanism of protein regulation by ubiquitylation. At the same time functional disturbances forestall further ubiquitin-dependent transformation of target protein (for example, targeting for proteosomal degradation). It is quite probable that influence of ubiquitylation on protein

functional activity is very important in terms of regulation and does not depend of degradation function, the more so as ubiquitylation is a reversible process and ubiquitin can be removed from target protein by specific deubiquitylating enzymes (DUBs).

Our findings based on computational analysis of 3D structures of selected target proteins well correspond with experimental data on modification of another ubiquitylation substrates – NO synthase and cytochrome P450. In particular, it was shown that inactive monomeric form of NO synthase undergoes ubiquitylation predominantly in comparison with active oligomer (Bender *et al.*, 2000; Dunbar *et al.*, 2004). For P450 co-localization of acceptor lysine residue and active site was also shown (Banerjee *et al.*, 2000). Another example is CDC48 – multifunctional ATPase, where ubiquitylated lysine is situated between AAA ATPase domains.

All together these data suggest biological relevance of proposed regulatory mechanism.

REFERENCES

Banerjee A. *et al.* (2000) Identification of a ubiquitination-target/substrate-interaction domain of cytochrome P-450 (CYP) 2E1. *Drug Metab. Dispos.*, **28**, 118–124.

Bender A.T. *et al.* (2000) Ubiquitination of neuronal nitric-oxide synthase *in vitro* and *in vivo*. *J. Biol. Chem.*, **275**, 17407–17411.

Dunbar A.Y. *et al.* (2004) Ubiquitination and degradation of neuronal nitric-oxide synthase *in vitro*: dimer stabilization protects the enzyme from proteolysis. *Mol. Pharmacol.*, **66**, 964–969.

Haglund K., Dikic I. (2005) Ubiquitylation and cell signaling. *EMBO J.*, **24**, 3353–3359.

Laub M. *et al.* (1998) Modulation of calmodulin function by ubiquitin-calmodulin ligase and identification of the responsible ubiquitylation site in vertebrate calmodulin. *Eur. J Biochem*, **255**, 422–431.

Peng J. *et al.* (2003) A proteomics approach to understanding protein ubiquitination. *Nat. Biotechnol.*, **21**, 921–926.

Shin H. *et al.* (2003) Identification of ubiquitination sites on the X-linked inhibitor of apoptosis protein. *Biochem. J.*, **373**, 965–971.

Smiley J.A., Jones M.E. (1992) A unique catalytic and inhibitor-binding role for Lys93 of yeast orotidylate decarboxylase. *Biochemistry*, **31**, 12162–12168.

PREDICTION IN CHANGES OF PROTEIN THERMODYNAMIC STABILITY UPON SINGLE MUTATIONS

***Demenkov P.S.*[*1, 2], *Ivanisenko V.A.*[2]**
[1] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: demps@math.nsc.ru

Key words: thermodynamic stability, single mutation, computer analysis

SUMMARY

Motivation: The creation of artificial proteins is a great challenge in today's biology. Prediction of the experimental results for changes in proteins surely can considerably accelerate the development of novel proteins.

Results: We have derived rules for the prediction of changes in protein thermodynamic stability upon introduction of single substitution in sequence. Using models of neural networks, backward propagation errors, and the modified KRAB method have established the rules. Based on the methods, we developed software allowing us to predict protein free energy upon single substitutions. In this work, we also compare the results. It was demonstrated that the modified KRAB algorithm, when based on the available data, allowed us to predict changes in thermodynamic stability with higher accuracy compared with back propagation networks.

INTRODUCTION

The creation of artificial mutant proteins with preassigned properties has been a challenge for biologists. Artificial mutations required the replacement of a particular amino acid by some other in the linear structure of protein (Afonnikov, 2002). Currently this is achieved experimentally: designers replace amino acids and make their inferences (Yanase *et al*., 2005; Canadillas *et al.*, 2006). High cost is the major drawback of this approach. There is an obvious need in reducing cost without efficiency loss. The first step is to develop mathematical methods, then to derive from them software able to predict the results of the designed experiments. It should be stipulated that, for a given protein, most appropriate mutation versions are sought, such that spare the overall structure of the protein without affecting or increasing its thermodynamic stability.

The Protein Data Bank (PDB) (Berman *et al.*, 2000) was used to tackle the problem of defining the amino acid environment at the location of interest in space. The other source for model building was the information about changes in protein free energy stored in the ProTherm database (Gromiha *et al.*, 2000). We were aware that free energy is an indicator of the changes in protein thermodynamic stability.

Using a neural network-based method and support vector machine allows to predict the sign of changes in protein free energy upon single point mutation (Capriotti *et al*., 2004, 2005). The accuracy of these predictors is high, approximately 75 %. In this paper, we present methods that enabled us to predict not only the positive or negative signs, but also the neutral changes in protein free energy upon single point mutation.

DATABASE PROCESSING

The functional and structural data for proteins are stored in PDB. We proceeded on the assumption that amino acids in the nearest environment of the spatial location of interest are consequential effect for the amino acid type at the location of interest. Based on the above assumption we developed a program that converted the PDB information into a table. Every protein location around which substitutions were to be replaced were encircled by spheres, r = 10Å. Then, the amino acid types within every sphere, the type of secondary amino acid structure at the examined location, and the relative solvent accessibility (RSA) values were stored. Information about RSA can be very useful. This is because mutations on protein surface affect much less free energy than those in the protein core (Guerois *et al.*, 2002). Information about amino acid types within the drawn sphere was encoded by a 20-dimensional vector. Every vector component was compared with a particular (1 of the 20) amino acid type, and the vector component was equal to the number of amino acid types within the drawn sphere. We considered four secondary structure types (Strand, Helix, Coil, Turn) defined by the algorithm STRIDE Frishman, Agros (1995). The secondary structure type was encoded in the same way using a vector of the 4-dimensional space. Each component of the vector corresponded to one of these 4 secondary structure types.

We took information about changes in the free energy ($\Delta\Delta G$) upon the introduction of the single mutation in the protein sequence from the ProTherm database. The information about the amino acid types before and after the introduced mutation was collected. Also, we collected information about the ambient experimental conditions (temperature, pH). Information about the mutations was coded by a 20-dimensional vector. These 20 elements code for the corresponding 20 amino acid residues, the element corresponding to the amino acid type before the mutation we defined as "-1", the element corresponding to the deleted residue, and as "1" the introduced residue (the one after mutation). All the remaining elements were kept equal to 0.

We considered two datasets. The S1 contained information about amino acids within the local environment of the amino acid around which the mutation was introduced; also, in addition the S2 contained information about the secondary structure type at the examined position.

Thus, S1 contained vectors of 43 components and S2 had 47 components. S1 and S2 contained 2126 vectors.

We did not envisage predicting the accurate value of changes in the free energy caused by the introduced mutation, rather we strived to define the direction of the change, i.e. our intention was to predict the variability or invariance of the protein free energy. Accordingly, their values ($\Delta\Delta G$) were assigned to groups:

1. $\Delta\Delta G > 0.1$
2. $-0.1 < \Delta\Delta G < 0.1$
3. $\Delta\Delta G < -0.1$

ALGORITHM DESCRIPTION BACK PROPAGATION NETWORK

To tackle the issue, two neural networks were used. Both were one-layer perceptron. The N1 had 43 input neurons, it contained 4 neurons in the hidden layer; its output had 3 neurons. The N1 was trained and tested on the S1 dataset. The resulting unity in the first output neuron as a result of network function increased ($\Delta\Delta G > 0.1$), it remained unaltered in the second ($-0.1 < \Delta\Delta G < 0.1$) and decreased in the third ($\Delta\Delta G < -0.1$).

The N2 structure contained 47 neurons in input. Four additional neurons were used to code the amino acid structure of the secondary type. It contained 6 neurons in the hidden layer; its output had 3 neurons. The N2 was trained and tested on the S2 dataset.

The value of 0.5 was accepted as the threshold for the output of the neural networks. When the value was greater than 0.5 at the network output, the value was accepted as unity (rounded up), when the output value was smaller than 0.5, the accepted value was 0 (rounded down).

If the outputs of the neural networks consisted of more than one unity, then our choice of one of the 3 ΔΔG value groups was based on the absolute values of outputs of the neural networks.

KRAB ALGORITHM

KRAB (Zagoruyko, 1999) was the second algorithm. Let vectors from S1 and S2 be points in multivariate space. The distance between points we define through the Euclidian's distance.

Work with KRAB started with the finding of a pair of points with the minimum distance between them. The found pair was connected with an edge. Then, the next nearest point pairs of those that were not as yet connected to the already built graph were linked. The procedure was reiterated until all the points were connected by edges. Such a graph is loopless, and the total length of all its joined edges is minimal. The graph which such features is known as the shortest open pathway (SOP) (Prim, 1961). To divide the graph into two parts, the longest edge linking the vertices of different types was removed. The operation was iterated until a subdivision was achieved that include graph points of only one type in each and every class. A class is a set of points connected by edges. The representative element whose characteristics are the averages for a class was compared with each class. The type of a vector is defined by type of the nearest representative element.

RESULTS

The datasets S1 and S2 were subdivided into two portions of the same size. Each portion contained **1063** vectors. One portion was used to train the algorithms, the other was applied to validate the accuracy of the established prediction rules. Accuracy was expressed as the percent of the correctly predicted values for the direction of changes in the free energy among the elements of the validated samples. The results are summarized in Table 1.

Clearly, the information about the secondary structure type proved to be useful in the prediction of changes in protein free energy. Using two neural networks and KRAB, we achieved an increase in prediction accuracy when relying on the S2 set. It is noteworthy that the KRAB algorithm provides improved accuracy compared to the neural networks we proposed. Also, training of the KRAB algorithm is less time consuming.

Table 1. Performance of the applied methods

Set Method	S1	S2
N1	68.08 %	–
N2	–	72.58 %
KRAB	73.41 %	75.83 %

ACKNOWLEDGEMENTS

Work was supported in part by Russian Foundation for Basic Research (Nos 04-01-00458, 05-04-49283, 06-04-49556), the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction" federal goal-oriented technical program "Study and

development of priority directions in science and technique, 2002–2006, Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science", the CRDF Rup2-2629-NO-04, Leading Science School (SS-4413.2006.1).

REFERENCES

Afonnikov D.A. (2002) *Computer analysis of coordinated changes in amino acid substitutions and families of homologous protein sequences*. Institute of Cytology and Genetics SB RAS, M.S. thesis. (In Russ.).

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.

Canadillas J.M., Tidow H., Freund S.M., Rutherford T.J., Ang H.C., Fersht A.R. (2006) Solution structure of p53 core domain: structural basis for its instability. *Proc. Natl. Acad. Sci. USA*, **103(7)**, 2109–2114.

Capriotti E., Fariselli P., Casadio R. (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. *Bioinformatics*, **20,** i63–i68.

Capriotti E., Fariselli P., Calabrese R., Casadio R. (2005) Predicting protein stability changes from sequences using support vector machines. *Bioinformatics*, **21,** Suppl 2, ii54–ii58.

Frishman G.E., Argos. (1995) Knowledge-Based Protein Secondary Structure Assignment. *Proteins*, **23,** 566–579.

Gromiha M.M., An J., Kono H. *et al*. (2000) Pro Therm, version 2.0:thermodynamic database for proteins and mutants. *Nucl. Acids Res*., **28**, 283–285.

Guerois R., Nielsen J.E., Serrano L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.,* **320**, 369–387.

Prim Z.L. (1961) The Shortest conjunctive networks and some generalizations. *Cybernetics collection*, **2,** 95–107 (In Russ.).

Yanase M, Takata H, Fujii K, Takaha T, Kuriki T. (2005) Cumulative effect of amino acid replacements results in enhanced thermostability of potato type L alpha-glucan phosphorylase. *Appl Environ Microbiol*., **71**(9), 5433–5439.

Zagoruyko N.G. (1999) *Applied methods for analysis of data and knowledge.* Publishing House of the Institute of Mathematics, Novosibirsk, Russia. (In Russ.).

EFFECT OF THE STRUCTURAL CONTEXT ON SPECIFICITY OF INTRA- AND INTERHELICAL INTERACTIONS IN PROTEINS

Efimov A.V.*, Brazhnikov E.V., Kondratova M.S.
Institute of Protein Research, RAS, Pushchino, Moscow Region, 142290, Russia
* Corresponding author: e-mail: efimov@protres.ru

Key words: α-helix packing, H-bond, rotamer, salt bridge, side chain

SUMMARY

Motivation: In proteins, α-helices can be packed in some different ways and each type of the α-helix packing forms a specific structural environment (or structural context) of side chains forming the interface.

Results: A stereochemical analysis of intra- and interhelical side chain–side chain and side chain–main chain interactions in different α-helical packings enable us to show that the specificity of these interactions is dependent not only on the physico-chemical character of residues but also on their structural context.

INTRODUCTION

Hydrogen bonding, ionic and hydrophobic interactions play important roles in stabilizing the native structure of a protein as well as in protein folding. It is widely believed that, whereas nonspecific hydrophobic interactions contribute to protein stability, the polar interactions can impart specificity to protein folding. An analysis of the frequency of occurrence of interhelical polar side chain–side chain pairs connected by hydrogen bonds or salt bridges in proteins shows that some of them occur frequently, others rarely, and there are those not to occur at all (Efimov, Kondratova, 2003). It is reasonable to assume that higher frequencies of occurrence of some interhelical pairs show that interactions are favorable for these side chain–side chain pairs compared to others. This does not mean that one side chain "recognizes" the other. The specificity appears at the level of higher order structures, for example, in pairs of closely packed α-helices.

METHODS AND ALGORITHMS

For this study, a data set of 120 non-homologous globular α-proteins and 45 coiled coils was used. Interhelical hydrogen bonds were determined using WHAT_IF (http://www.cmbi.kun.nl:1100/WIWWWI/). Interhelical salt bridges were determined with the use of our own software.

IMPLEMENTATION AND RESULTS

In proteins, α-helices can be packed in some different ways and each type of the α-helix packing forms a different structural environment (or structural context) of side chains forming the interface and taking part in interactions. There are two main ways that amphipathic α-helices pack against each other. In the first case, two α-helices are packed so that their hydrophobic side chains form a double layer in the packing interface. Here, hydrophobic stripes of the α-helices interact in a face-to-face manner and hence this is referred to as the face-to-face packing of α-helices. In the case of a side-by-side packing of α-helices, their hydrophobic stripes associate in a side-by-side manner and form a common hydrophobic surface on the bihelical structure. In each case, α-helices can be packed either parallel or antiparallel. Two α-helices neighboring in the chain, packed side-by-side and antiparallel can form either a right-turned or left-turned α-α-hairpin (for details, see Efimov, 1979, 1999).

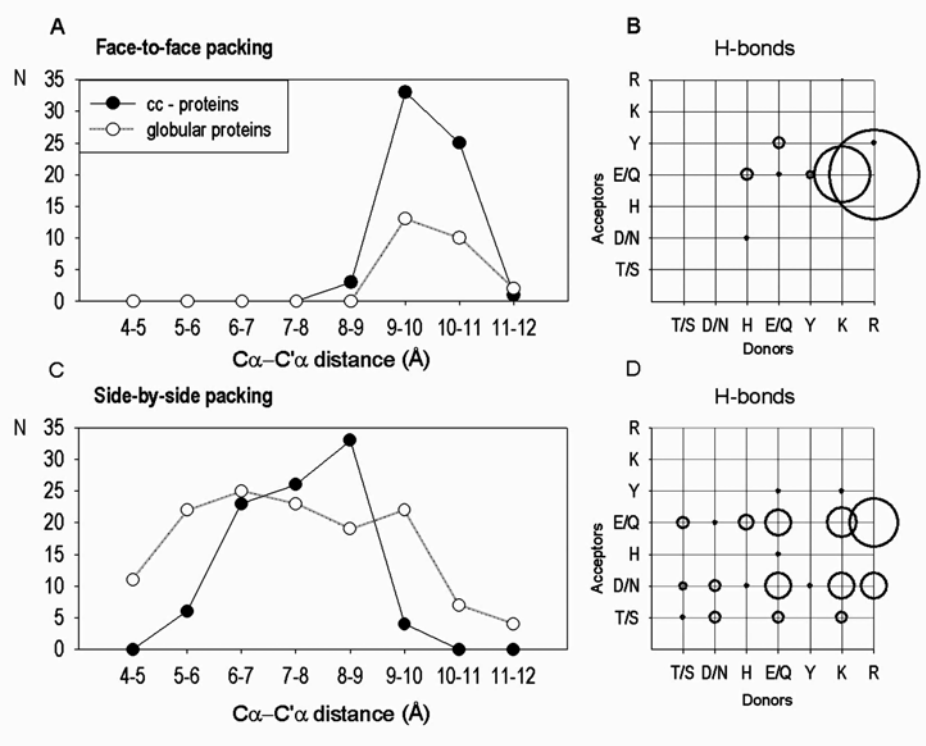


Figure 1. Comparison of distances between backbone surfaces forming the interface of pairs of α-helices packed in α face-to-face (*A*) or side-by-side manner (*C*). The C_α-$C_{\alpha'}$ distance (the prime denotes belonging to another helix) is an average distance between C_α-atoms of closest residues forming the interface. N is the number of α-helical pairs having the corresponding distance. Circular diagrams on the right show the frequency of occurrence of sidechain-sidechain pairs forming interhelical H-bonds in the corresponding sets of α-helical pairs. Donors and positively charged groups of one helix are laid of on the horizontal axes, while ordinates show acceptors and negatively charged groups of the other helix. The radius of the circle is directly proportional to the frequency of occurrence of the corresponding sidechain-sidechain pair.

A stereochemical analysis of these α-helical packings in a data set of 120 non-homologous globular α-proteins and 45 coiled coils has shown that:

i) On average, backbone surfaces forming the interface are arranged closer in the side-by-side packings than in face-to-face packings of α-helices (Fig. 1*A*, *C*).

ii) In pairs of α-helices packed face-to-face, the interhelical H-bonds and salt bridges are formed, as a rule, between long side chains (most often Lys-Glu, Lys-Gln, Arg-Glu

and Arg-Gln pairs, see Fig. 1*B*), and those in the side-by-side packings are formed by both long and short side chains (Fig. 1*D*). This appears to be one of the most important determinants of specificity of the α-helix packing in proteins. For example, if two interacting α-helices have no long side chains in the corresponding positions, they can not be packed face-to-face but can be packed side-by-side.

iii) Each type of the α-helix packing has its specific set of rotamers of hydrophobic side chains in a- and d-positions. In other words, selection of side chain rotamers in a- and d-positions of α-helices depends on the type of the α-helix packing and consequently on the structural context. In order to demonstrate this feature we used a representative set of 13 coiled-coil dimers in which α-helices are packed face-to-face and parallel: 1D7M, 1DH3, 1GD2, 1ZII, 1UIX, 1KDD, 1CZ7, 1KQL, 1S9K, 1LLM, 1CE9, 1T6F, 1P9I. Fig. 2 shows that in these proteins most side chains of leucines found in a-positions have *trans*-isomers ($\chi_1 \cong 180°$) and those in d-position – -*gauche*-isomers ($\chi_1 \cong -60°$). This is a characteristic of α-helices packed face-to-face and parallel. Other packings of α-helices have different sets of side-chain rotamers (Efimov, 1979; Brazhnikov, Efimov, 2006, in preparation). For example, in α-helices packed side-by-side and parallel (as found, e.g., in coiled-coil tetramers) most of Leu residues occupied a-positions have side-chain g-rotamers and those in d-positions t-rotamers. It should be noted that these strong rotamer preferences depending on the residue position and the structural context have been demonstrated for the first time. Earlier computational studies have described the identification and classification of side-chain rotamers as well as the frequency of occurrence of different rotamers depending on the local secondary structure (see, e.g., Dunbrack, Cohen, 1997; Lovell *et al*., 2000).

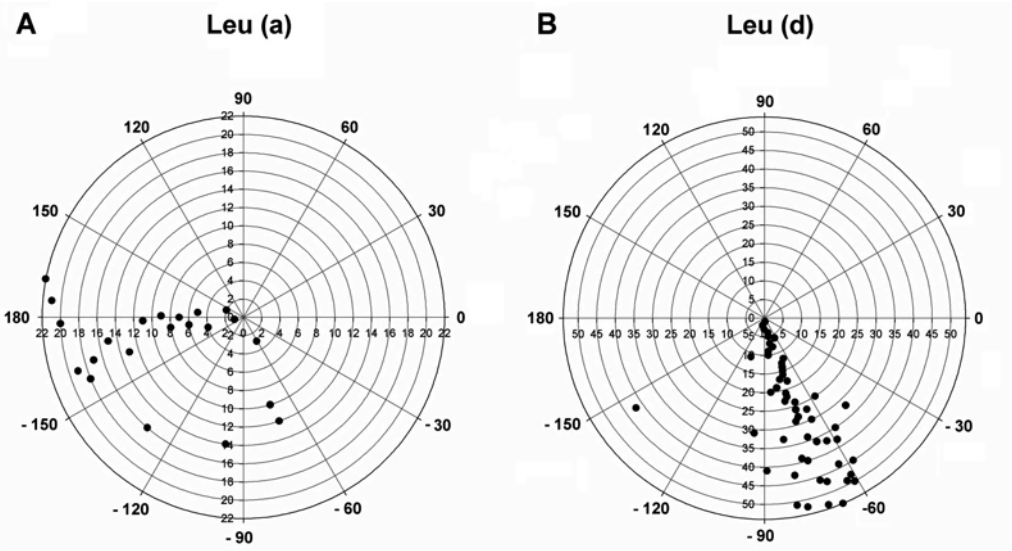


Figure 2. Distribution of torsion angles of leucine side chains found in a-positions (*A*) and d-positions (*B*) of 13 coiled-coil dimers.

DISCUSSION

A principal prerequisite of bonding is known to be certain proximity of the partners to each other. For H-bonding, the donor-acceptor distance should be less than 3,5 Å (according to WHAT_IF criteria) and in salt bridges the distance between heavy atoms of oppositely charged groups should be less than 4 Å. In face-to-face packings of α-helices, bulky hydrophobic side chains are located in the interface and this results in larger interhelical distances as compared with side-by-side packings, where hydrophobic side chains are

arranged on the surface. In our opinion, this is the main reason that only long polar side chains form interhelical H-bonds and salt bridges in face-to-face packings of α-helices (Fig. 1*A*, *B*). In side-by-side packings, the interhelical distance is such that short side chains are able to form interhelical H-bonds (Fig. 1*C*, *D*). On the other hand, intra- and interhelical interactions between hydrophobic side chains also differ in face-to-face and side-by-side packings of α-helices thus resulting in different sets of hydrophobic side chain rotamers.

ACKNOWLEDGEMENTS

This work was supported in part by the Russian Foundation for Basic Research (Grant No. 04-04-49343a).

REFERENCES

Dunbrack R.L., Cohen F.E. (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. *Protein Science*, **6**, 1661–1681.

Efimov A.V. (1979) Packing of α-helices in globular proteins. Layer-structure of globin hydrophobic cores. *J. Mol. Biol.,* **134**, 23–40.

Efimov A.V. (1999) Complementary packing of α-helices in proteins. *FEBS Lett.,* **463**, 3–6.

Efimov A.V., Kondratova M.S. (2003) A comparative analysis of interhelical polar interactions of various α-helix packings in proteins. *Mol. Biol., (Mosc.),* **37**, 515–521.

Lovell S.C. *et al.* (2000) The penultimate rotamer library. *Proteins,* **40**, 389–408.

RISE OF NEW Zn^{2+} BINDING SITES CAN BE A MOLECULAR MECHANISM FOR IMPAIRED FUNCTION OF THE p53 MUTANTS

***Fomin E.S.*[1], *Oshurkov I.S.*[2], *Ivanisenko V.A.*[*1, 2]**
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia
[*] Corresponding author: e-mail: salix@bionet.nsc.ru

Key words: p53, mutations, rise of new binding site, molecular mechanism of impaired function, molecular simulation

SUMMARY

Motivation: At present, there are abundant experimental data concerning associations between the mutations in the p53 protein and cancer. Not of all the molecular mechanisms underlying the impairment of p53 function are known.

Results: We support here our previous assumption that G245C mutations can give rise to an additional Zn binding site in the immediate vicinity to the functionally significant binding site (Ivanisenko *et al*., 2005). We demonstrated here that the interaction energy of the Zn ion in the G245C mutant with the *de novo* arisen site is commensurate with the interaction energy in the wild-type p53. The presence of an additional site in the mutant can damage p53 conformation upon interaction with DNA. Also, using molecular mechanics, we calculated the effects of certain other mutations on the zinc interaction energy with the normal site.

INTRODUCTION

The p53 tumor suppressor is a transcription factor. In response to various types of genotoxic stresses, p53 transactivates a number of genes by binding to specific DNA sequences (el-Deiry *et al*., 1992), thereby arresting cell cycle, repairing damaged DNA, or inducing apoptosis as the cell fates (Giaccia, Kastan, 1998; Jin, Levine, 2001). The structure of the p53 core DNA-binding domain (residues 94–312) that binds directly to the DNA sequence has been resolved by X-ray crystallography. The resolved structures have been obtained for certain p53 mutations, for example, the crystal structure of a super stable mutant of human p53 core domain (Joerger *et al*., 2004).

In about half of all the human cancers, p53 is inactivated as a direct result of missense mutations within the p53 gene. Most of these mutations map to the DNA-binding core domain (Hainaut, Hollstein, 2000), and six "hot spots" stand out as the most frequently associated with human cancer (R175H, G245S, R248Q, R249S, R273H, and R282W). However, little is known about the underlying molecular mechanism of impaired p53 function upon mutations. There exist mechanisms implying how the tumorigenic p53 core domain mutations can truly cause reduction in p53 site-specific DNA binding activity (Kern *et al*., 1992; Martin *et al*., 2002). The putative mechanisms assume the eliminating critical protein-DNA contact like R273H (Bullock *et al*., 1997; Wong *et al*., 1999), lowering thermodynamic stability, like I195T (Friedler *et al*., 2003), or enhancing loss of the single bound Zn^{2+} ion, like R175H (Butler, Loh, 2003). The Zn-free p53 core domain

appears to promote aggregation of Zn-bound p53 via a nucleation-growth process. Through a combination of induced p53 aggregation and diminished site specific DNA binding activity, Zn^{2+} loss may represent a significant inactivation pathway for p53 in the cell (Butler, Loh, 2003).

Our relevant assumption was that certain mutations can give rise to novel functional sites (for details, see Ivanisenko *et al*., 2005). We found an extra Zn binding site that overlaps the normal Zn binding site in the mutant protein, G245C. According to the X-ray crystal structure of the p53 core domain (PDB ID 1gzh), Zn^{2+} is coordinated to the C176, H179, C238, and C242 residues. The mutation G245C gives rise to a new site (H179, C242, and C245) similar in structure to that for Zn^{2+} cystidine deaminase binding [PDB ID 1af2, (Ivanisenko *et al*., 2005)]. We report here the energies calculated for the interaction of the Zn^{2+} ion with the wild-type and the novel site. The data were obtained using molecular mechanics. It was demonstrated that the calculated energies agree with each other. This is evidence that supports the functionality of the novel site.

METHODS AND ALGORITHMS

The spatial structures of the human p53 core domain were used. These included the wild-type structure (PDB ID 1gzh), the model structure of the R175H mutant obtained by the SCWRL3.0 program (Canutescu *et al*., 2003), also the model structure of the G245C mutant we previously used (Ivanisenko *et al*., 2005). The initial position of Zn^{2+} for the Zn-protein complex in the case when Zn is bound to the new site in the G245C mutant was defined using the PDBSiteScan program (Ivanisenko *et al*., 2004). Unlimited geometry optimization of the Zn - protein complex in the area of its binding site was used. The L-BFGS energy minimization method implemented in the GROMACS 3.3.1 package (Berendsen *et al*., 1995; Lindahl *et al*., 2001) was applied. The minimization convergence limit 5 kJ/mol*nm was used. To maintain electroneutrality, the appropriate amounts of Cl^- anions were added, the water molecules were presented as point charges (SPC-model). The OPLS-AA/L all-atom force field was used; the cut-off value for the van-der-Waals potential was 1 nm; the Coulomb interaction was calculated by the PME method.

Initial calculations demonstrated that, depending on the starting geometry and the chosen optimization method, the final configurations of the system differ by the interaction energy of Zn^{2+} with the protein (different local minima) in the 10–20 kJ/mol range, and, in certain cases, up to 50 kJ/mol. To minimize the error due to the "escape" of the system to acquire novel unsuitable configurations, we used the same calculation procedure for all the mutant proteins. The procedure was as follows. (1) a single geometry of wild protein was accepted as the starting geometry, mutations were induced in the geometry using the SCREWL3.0 program that does not affect the nonmutated amino acid residues; (2) the binding site geometry was partly "frozen" (Zn ion and the its four nearest atoms) and the entire system with convergent limit 10 kJ/mol*nm was minimized; (3) the binding site atoms were "unfrozen", and the minimization procedure with convergent limit 5 kJ/mol*nm was reiterated; it should be noted that the increase in accuracy from 10 to 5 kJ/mol*nm produces the correction in interaction energy between Zn^{2+} and protein in 1–3 kJ/mol range; (4) finally, water molecules and chloride anions were added, and the minimization process was reiterated.

The GROMACS 3.3.1 program did not allow us define directly the deprotonized charge state for cystein. To define the required charge state, we modified the *.rtp files. We added a description for cystein in the deprotonized charge state. The charges on the atoms (according to Mulliken) were calculated using the semiempirical MINDO/3 method (Dewar, 1975). The calculation parameters were as follows. The total charge was -1, spin multiplicity was 1. Unlimited Hartree-Fock calculation was performed.

We did not calculate the binding energy for Zn^{2+} because it is expressed as the difference between two large values, the total energies of the entire system with Zn^{2+} and without it. This produces a serious error. Instead, we calculated Zn^{2+} – protein interaction that includes Coulomb and van-der-Waals terms. We assume that interaction and binding energies differ by a constant.

RESULTS AND DISCUSSION

To validate the accuracy of our approach, we calculated the interaction energy of Zn^{2+} with the R175H mutant, which loses zinc at physiological temperature (Butler, Loh, 2003). The table gives the calculated interaction energies $E_{total} = E_{Coulomb} + E_{van\text{-}der\text{-}Waals}$ of Zn^{2+} with wild-type protein and mutants in the deprotonized cystein charge states of protein.

Table 1. The interaction energy between Zn^{2+} with wild-type and mutant protein

Name	$E_{Coulomb}$(kJ/mol)	$E_{van\text{-}der\text{-}Waals}$(kJ/mol)	E_{total} (kJ/mol)
Wild	-1328.99	136.61	-1192.38
$R175H_1$	-1319.38	137.14	-1182.24
$G245C_2$	-1333.24	139.99	-1193.25
$G245C^2$	-1338.12	143.14	-1194.98

[1]– Zinc is placed at the position where it interacts with the wild-type site.
[2]– Zinc is placed at the position where it interacts with the novel site.

As expected for the R175H mutant, the interaction energy for the Zn^{2+} was weaker. This is consistent with the experimental data for zinc loss upon this mutation.

As the table shows, the interaction energies of Zn^{2+} with wild-type site and novel site for the G245C mutant are comparable in their values. This supports our idea that the novel site for the G245C mutant is functionally competent. Impaired function of the p53 G245C mutant may be a result of the competition of these sites for Zn^{2+}.

The mutation G245C was found in families with the Li-Fraumeni syndrome (Malkin *et al*., 1990). From analysis of the functional significance of this germline mutation, it was concluded that malignant cells lose tumor-suppressor activity (Frebourg, 1996). Our results suggest a molecular mechanism for the effect of the G245C substitution based on competition between normal and extra sites for Zn^{2+} binding.

It should be noted that our approach to the modeling of the interaction of the metal ion – protein complex using the GROMACS package is the first approximation. Approaches relying on combined quantum chemistry with molecular mechanics such as QM-MM would hopefully provide further insights.

ACKNOWLEDGEMENTS

Work was supported in part by Russian Foundation for Basic Research No. 05-04-49283, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation “Identification of promising targets for the action of new drugs on the basis of gene network reconstruction” federal goal-oriented technical program “Study and development of priority directions in science and technique, 2002–2006”, Interdisciplinary integrative project for basic research of the SB RAS N115 “Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science” and the CRDF Rup2-2629-NO-04.

REFERENCES

Berendsen H.J.C. *et al*. (1995) GROMACS: A message-passing parallel molecular dynamics implementation, *Comp. Phys. Comm.*, **91**, 43–56.

Bullock A.N. *et al.* (1997) Thermodynamic stability of wild-type and mutant p53 core domain. *Proc. Natl. Acad. Sci. USA*, **94**, 14338–14342.

Butler J.S, Loh S.N. (2003) Structure, function, and aggregation of the zinc-free form of the p53 DNA binding domain. *Biochemistry*, **42**, 2396–2403.

Canutescu A.A. *et al.* (2003) A graph theory algorithm for protein side-chain prediction. *Protein Science,* **12**, 2001–2014.

el-Deiry *et al*. (1992) Definition of a consensus binding site for p53. *Nat. Genet.*, **1**, 45–49.

Dewar M.J.S. *et al*. (1975) *QCPE*, **11**, 279.

Frebourg T. *et al*. (1992) Germ-line mutations of the p53 tumor suppressor gene in patients with high risk for cancer inactivate the p53 protein. *Proc. Nat. Acad. Sci. USA*, **89**, 6413–6417.

Friedler A. *et al.* (2003) Kinetic instability of p53 core domain mutants: implications for rescue by small molecules. *J. Biol. Chem.*, **278**, 24108–24112.

Giaccia A.J., Kastan, M. B. (1998) The complexity of p53 modulation: emerging patterns from divergent signals. *Genes Dev*., **12**, 2973–2983.

Hainaut P., Hollstein M. (2000) *Adv. Cancer Res*., **77**, 81–137.

Ivanisenko V.A *et al*. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucl. Acids Res.*, **32**, W549–W554.

Ivanisenko V.A. *et al*. (2005) PDBSite: a database of the 3D structure of protein functional sites. *Nucl. Acids Res*., **33**, D183–187.

Jin S., Levine A.J. (2001) The p53 functional circuit. *J. Cell Sci.*, **114**, 4139–4120.

Joerger A. *et al*. (2004) Crystal structure of a superstable mutant of human p53 core domain. Insights into the mechanism of rescuing oncogenic mutations. *J. Biol. Chem.*, **279**, 1291–1296.

Kern S.E. *et al*. (1992) Oncogenic forms of p53 inhibit p53-regulated gene expression. *Science*, **256**, 827–830.

Lindahl E. *et al.* (2001) GROMACS 3.0:A package for molecular simulation and trajectory analysis. *J. Mol. Mod*., **7,** 306–317.

Malkin D. *et al.* (1990) Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science*, **250**, 1233–1238.

Martin A.C. *et al*. (2002) Integrating mutation data and structural analysis of the TP53 tumor-suppressor protein. *Hum Mutat*., **19**, 149–164.

Wong K.B. *et al.* (1999) Hot-spot mutants of p53 core domain evince characteristic local structural changes. *Proc. Natl. Acad. Sci. USA*, **96**, 8438–8442.

SEQUENCE-BASED PREDICTION OF DNA-BINDING SITES ON DNA-BINDING PROTEINS

Gou Z., Hwang S., Kuznetsov B.I.*
Gen*NY*sis Center for Excellence in Cancer Genomics, University at Albany,
One Discovery Drive, Rensselaer, NY, USA
* Corresponding author: e-mail: ikuznetsov@albany.edu

Key words: protein-DNA interaction, position specific scoring matrix, evolutionary conservation, web-server, DNA binding, prediction, pattern recognition, machine learning

SUMMARY

Motivation: Identification of DNA-binding sites on DNA-binding proteins is important for functional annotation. Experimental determination of the structure of a protein-DNA complex is an expensive process. Reliable computational methods that utilize the sequence of a DNA-binding protein to predict its DNA-binding interface are needed.

Results: We present an application of three machine learning methods: support vector machine, kernel logistic regression, and penalized logistic regression to predict DNA-binding sites on a DNA-binding protein using its amino acid sequence as an input. Prediction is performed using either single sequence or a profile of evolutionary conservation. The performance of our predictors is better than that of other existing sequence-based methods. The outputs of all three individual methods are combined to obtain a consensus prediction. This further improves performance and results in accuracy of 82.4 %, sensitivity of 84.9 % and specificity of 83.1 % for the strict consensus prediction.

Availability: http://lcg.rit.albany.edu/dp-bind.

INTRODUCTION

A reliable identification of DNA-binding sites on DNA-binding proteins is important for *in silico* modeling of protein-DNA interactions and functional annotation. Identification of DNA-binding sites is relatively straightforward if the structure of a protein-DNA complex is known. However, solving the structure of a protein-DNA complex is a very complicated and time-consuming process. Several computational methods that use experimentally solved unbound structure of a DNA-binding protein to identify DNA-binding interface based on the electrostatic potential and the shape of molecular surface have been developed (Jones *et al*., 2003; Tsuchiya *et al*., 2004). However, these methods cannot be used if experimentally determined protein structure is not available. An alternative to the structure-based prediction is a sequence-based prediction. In this work, we apply a combination of three supervised pattern recognition methods to improve the prediction of DNA-binding sites in a DNA-binding protein using its amino acid sequence as the only input.

METHODS AND ALGORITHMS

Dataset of protein-DNA complexes. We used a non-redundant set of 62 experimentally solved protein-DNA complexes that were utilized previously to develop DBS-PRED (Ahmad *et al*., 2004) and DBS-PSSM (Ahmad, Sarai, 2005). We label an amino acid residue in a protein chain as DNA-binding if the distance from at least one of its heavy atoms (atoms other than hydrogen) to any heavy atom in DNA is shorter than the cutoff distance of 4.5Å. In order to balance the number of examples between binding and non-binding residues, for each protein chain we randomly sampled without replacement the same number of non-binding residues as that of the DNA-binding ones.

SEQUENCE ENCODING

In order to represent the input protein sequence by a numerical feature vector, we used two types of sequence-based encoding and encoding based on PSI-BLAST (Altschul *et al*., 1997) position specific scoring matrix (PSSM). In the first type of sequence encoding, called binary encoding, the 20 amino acid types are represented by 20 mutually orthogonal binary vectors of dimension 20 (Qian, Sejnowski, 1988). In the second type of sequence encoding, called BLOSUM62 encoding, each amino acid type is represented by a vector of dimension 20 using a corresponding row from the BLOSUM62 amino acid substitution matrix (Henikoff, Henikoff, 1992). In the case of PSSM-based encoding, each sequence position is encoded by a 20-dimensional vector obtained from a corresponding row in the PSSM (Ahmad, Sarai, 2005). In both the BLOSUM62 and PSSM encoding, we normalize all elements in the matrix between 0 and 1 using the logistic function $f(x) = 1/[1+\exp(-x)]$. In all three encoding methods, nearest sequential neighbors of a sequence position are encoded with a standard procedure (Qian, Sejnowski, 1988) using a sliding window of size 7.

MACHINE LEARNING ALGORITHMS

For our two-class (DNA-binding and non-binding residues) classification problem, we applied three machine learning algorithms: support vector machine (SVM) (Christianini, Shawe-Taylor, 2000), kernel logistic regression (KLR) (Zhu, Hastie, 2005), and penalized logistic regression (PLR) (le Cessie, van Houwelingen, 1992). SVM is a margin maximizing classifier that does a linear classification in the feature space, which corresponds to a non-linear classification in the original data space. The feature space is obtained by transforming data from the original data space with a kernel function. Similarly, KLR and PLR are also margin maximizing classifiers. For both SVM and KLR we used the Radial Basis Function (RBF) kernel. The SVM algorithm was implemented using the LIBSVM program (http://www.csie.ntu.edu.tw/~cjlin/libsvm). We implemented the KLR and PLR algorithms in C++.

CONSENSUS PREDICTION

Each of the three machine learning methods independently assigns a label (binding or non-binding) to each position in the input sequence. Then, these three labels can be used to produce a consensus prediction for each sequence position. We used two types of consensus. The first is majority consensus obtained by majority voting (at least two of

three labels are identical). The other is strict consensus which retains only positions with high-confidence predictions on which all three methods agree.

EVALUATION OF THE PREDICTORS

We used leave-one-out cross-validation to train and test each predictor. We used accuracy (ACC), sensitivity (SN), and specificity (SP) to assess the performance of each predictor:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}, \quad SN = \frac{TP}{TP + FN}, \quad SP = \frac{TN}{FP + TN}$$

where TP, FN, TN and FP is the number of true positives (correctly predicted binding residues), false negatives (binding residues predicted as non-binding), true negatives (correctly predicted non-binding residues), and false positives (non-binding residues predicted as binding), respectively.

RESULTS AND DISCUSSION

ACC, SN, and SP of the predictors are shown in Table 1. Fig. 1 shows the receiver operating characteristics (ROC) curve for each predictor. ROC curve is more informative than most other measures and allows one to compare the performance of different classifiers by looking at the curve and the area under the curve (AUC). Larger AUC indicates better performance. Analysis of the data presented in Table 1 and Fig. 1 leads to the following observations:

1. All three individual sequence-based predictors have similar performance.
2. All three individual PSSM-based predictors have a significantly better performance than the sequence-based ones, PSSM-based KLR having the highest classification accuracy of 79.2 %.
3. The performance of PSSM-based KLR predictor (ACC of 79.2 %, SN of 76.4 %, SP of 82.0 %) is better than that of the other existing PSSM-based method for predicting DNA-binding sites, DBS-PSSM (ACC of 66.4 %, SN of 68.2 %, SP of 66.0 %).
4. The strict consensus prediction improves both sequence-based and PSSM-based predictions. The majority consensus performs better than individual methods in the case of single sequence-based prediction when evolutionary information is not utilized. It also improves sensitivity of the PSSM-based prediction.

A web server implementation of the predictors, called DP-BIND, is freely available at http://lcg.rit.albany.edu/dp-bind. It can be used for a high-confidence prediction of DNA-binding sites in a DNA-binding protein when its experimentally solved structure is not available.

Table 1. Measures of the performance of the predictors of DNA-binding sites (in percentage)

Classifiers	Sequence-based BLOSUM62 encoding			PSSM-based encoding		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
SVM	69.7±9.3	70.2±16.8	69.2±13.7	78.9±10.1	76.9±18.5	80.9±13.6
KLR	68.9±7.9	66.7±15.4	71.0±11.7	**79.2±10.0**	**76.4±18.5**	**82.0±12.4**
PLR	68.6±8.0	68.9±13.1	68.3±13.4	73.7±8.6	73.1±18.6	74.2±12.9
Majority consensus	70.5±8.8	71.3±9.8	72.0±11.4	78.9±10.1	81.3±10.5	80.1±13.4
Strict consensus	73.0±9.4	73.4±10.7	74.8±13.3	**82.4±10.8**	**84.9±11.0**	**83.1±13.3**

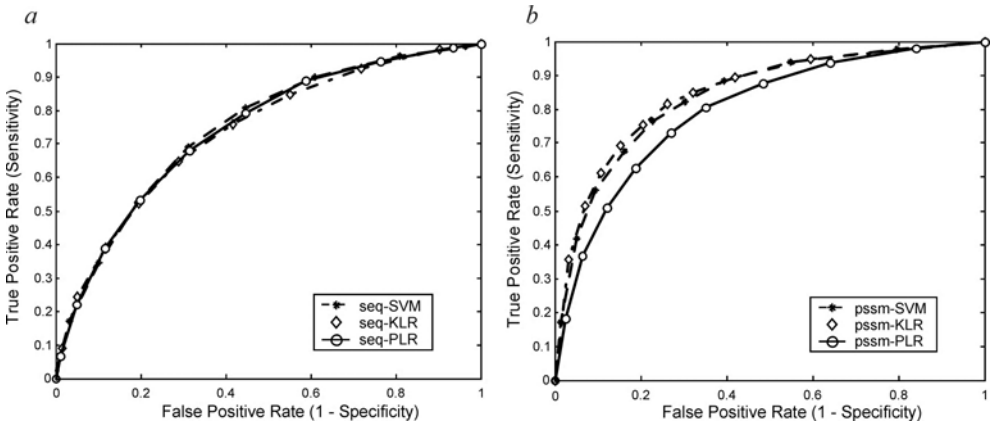


Figure 1. Receiver operating characteristics (ROC) curves for predictors that use (*a*) BLOSUM62 sequence-encoding and (*b*) PSSM-based encoding.

REFERENCES

Ahmad S. *et al*. (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, **20**, 477–486.

Ahmad S., Sarai A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, **6**, 33–38.

Altschul S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new gene ration of protein database search programs. *Nucl. Acids Res.*, **25**, 3389–3402.

Christianini N., Shawe-Taylor J. (2000) Support vector machines and other kernel-based learning methods. Cambridge University Press. Cambridge, MA.

Henikoff S., Henikoff J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

Jones S. *et al*. (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. *Nucl. Acids Res.*, **31**, 7189–7198.

le Cessie S., van Houwelingen J.C. (1992) Ridge estimators in logistic regression. *Appl. Statist.*, **41**, 191–201.

Qian N., Sejnowski T.J. (1988) Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.*, **202**, 865–884.

Tsuchiya Y. *et al*. (2004) Structure-based prediction of DNA-binding sites on proteins using the empirical preference of electrostatic potential and the shape of molecular surfaces. *Proteins*, **55**, 885–894.

Zhu J., Hastie T. (2005) Kernel logistic regression and the import vector machine. *J. Comp. Graph Stat.*, **14**, 185–205.

PDBSite DATABASE AND PDBSiteScan TOOL: RECOGNITION OF FUNCTIONAL SITES IN PROTEIN 3d STRUCTURE AND TEMPLATE-BASED DOCKING

***Ivanisenko V.A.*[*1, 2], *Ivanisenko T.V.*[3], *Sharonova I.V.*[2], *Krestyanova M.A.*[2], *Ivanisenko N.V.*[2], *Grigorovich D.A.*[1]**

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Novosibirsk State University, Novosibirsk, 630090, Russia; [3] Siberian State University of Telecommunications and Information Sciences, Novosibirsk, Russia

[*] Corresponding author: e-mail: salix@bionet.nsc.ru

Key words: PDBSite, PDBSiteScan, functional site, template-based docking

SUMMARY

Motivation: Recognition of functional sites in proteins is a direct computational approach providing a better understanding of protein biological and biochemical functions. Use of information about the protein spatial structure broadens our understanding of the structural organization of the functional sites, providing their recognition in the most efficient and accurate manner.

Results: We developed a new version for the PDBSite database that contains 3d templates of various protein functional sites (posttranslational modification, catalytic active, organic and inorganic ligand binding, protein-protein, protein-DNA and protein-RNA interactions) and also a new version of the PDBSiteScan tool ensuring the recognition of functional sites using 3d templates and the creation of molecular protein-ligand complexes relying on template based docking. The number of functional and drug binding sites stored in PDBSite was considerably increased, also, PDBSite was integrated with the other established molecular-biological databases.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsite/, http://wwwmgs.bionet.nsc.ru/mgs/systems/fastprot/pdbsitescan.html.

INTRODUCTION

Data on the protein 3d structure offer unprecedented advantages for studying the structural features of functional sites, the molecular mechanism underlying their function and, what is more, lends credibility to functional site predictions. Recently, many software tools have been developed for predicting protein functional sites. The tools aid functional site recognition in protein 3d structure based on computed chemical properties (Shehadi *et al*., 2005), as well as on databases containing structural data on active site and protein-ligand interactions (Laskowski *et al*., 2005a; Hendlich, 2003). There is now a repertoire of tools for the search of functional sites based on the detection of structural similarity to related proteins of known functions (Jones *et al*., 2003). Approaches to protein function prediction that integrate methods for the recognition of functional sites both in the spatial and primary protein structures, also genomic analysis (Laskowski *et al*., 2005b; Ko *et al*., 2005) are gaining deserved popularity. We have developed a database for the spatial structures of the protein functional sites, PDBSite (Ivanisenko *et al*., 2005). The PDBSiteScan program (Ivanisenko *et al*., 2004) uses the stored data to predict

posttranslational modification, catalytic active, organic and inorganic ligand binding, protein-protein, protein-DNA and protein-RNA interaction sites. Here, we implement novel, improved versions of PDBSite and PDBSiteScan.

METHODS AND ALGORITHMS

A brief description of the PDBSite structure and the PDBSiteScan program follows (for details, see Ivanisenko *et al*., 2004, 2005). PDBSite contains comprehensive structural and functional information on catalytically active centers of various enzymes, the sites of posttranslational protein modification, ion metal binding, binding organic/inorganic compounds, drug binding, protein-protein, protein-DNA and protein-RNA interactions. The data were extracted from the PDB databank on the basis of information in the SITE field of PDB indicating the amino acid residues of the functional sites; the sites of protein-protein, protein-DNA and protein-RNA interactions were identified by analysis of the atomic coordinates in their heterocomplexes. The sites included the amino acid residues that are in contact with the ligand (protein, RNA or DNA). The coordinates of the N, C-alpha and C-atoms of the functional sites from the PDBSite database are utilized by the PDBSiteScan program as site templates for the recognition of the functional sites. The new PDBSiteScan version provides the derivation of protein-ligand complexes from template based docking. To implement docking, we developed an auxiliary database known as the PDBLigand library. The PDBLigand library contains atom coordinates of the low molecular ligands, proteins, DNA and RNA, which bind to the sites from PDBSite. Template based docking is done by transfer of the ligand together with the site-template during the structural alignment of the site-template to protein. The generated draft protein-ligand complex can be accepted as an approximation to the further more accurate docking or molecular dynamics analysis.

RESULTS AND DISCUSSION

The PDBSite database description. The PDBSite database consists of structural and functional information about various protein functional sites. A database entry contains data about a single functional site that occurs in a particular protein. The structural data from a single entry of the database serve as template for the prediction of the site in the protein spatial structure by the PDBSiteScan program. According to the Gene Ontology the PDBSite database contains the functional sites for proteins that are classified into 951 various functions, 598 biological processes, and 165 cellular components. The total number of sites are distributed among the organisms as follows: Eukaryota – 13385, Bacteria – 7824, Archaea – 1743, Viruses – 1041. The templates are grouped into families according to the functional properties of the sites (see Table 1). Templates for the active sites are subdivided into 2 classes. The first class is composed of templates whose active site is not occupied by ligand and we called it “active, no ligand present”. The other class consists of active site complexed with ligand (substrate, metal ion, inhibitor and any other molecules) in the protein tertiary structure represented in the PDB database. We called it “active, ligand present”. The binding sites were divided also into 2 large classes: “single ligand present” and “multiple ligand present”. The multiple ligand present contains all the templates when several ligands are in contact with the site.

Table 1. Families of functional site templates and template number for each family

Active site[1]		Alkali Earth		Inorganic, organic and biochemical compounds	
Oxidoreductases	129/150	Sodium	188	950 families	5919
Transferases	115/105	Potassium	87	Multiple ligand present	
Hydrolases	676/459	Rubidium	5	300 families	2288
Lyases	103/150	Rare Earth		Drug binding	
Isomerases	38/46	Cesium	14	56 families	678
Ligases	6/3	Holmium	8	Protein-protein interaction	
Transition metal binding		Samarium	4	8000 families	10934
Zinc	724	Cerium	2	Protein-DNA interaction	
Copper	272	Europium	1	1500 families	2018
Manganese	220	Lutetium	1	Protein-RNA interaction	
Iron	144	Other Metals		1678 families	2492
Cadmium	122	Thallium	12	Posttranslational modification	
Nickel	67	Lead	5	Glycosylation	52
Cobalt	49	Gallium	1	Phosphorylation	24
Mercury	25	Non metal binding		Myristylation	5
Platinum	8	Chloride	173	Lipoylation	2
Yttrium	2	Iodide	9	Cleavage	5
Gold	1	Sulfur	3	Miscellaneous	4367
Alkaline Earth		Clusters			
Calcium	1059	Iron/Sulfur	159		
Magnesium	224	Iron/Sulfur/Oxygen hybrid cluster	7		
Strontium	1	Iron-MO-Sulfur	4		
		HF OXO cluster HF5	5		
		HF-OXO-Phosphate cluster HF3	1		
		HF-OXO-Phosphate cluster PHF	3		

[1] The number of templates is denoted as no ligand present/ligand present.

The PDBSite database is integrated with the Gene Ontology, UNIPROT, EMBL, PIR, TRANSFAC, ENSEMBL, INTERPRO, PFAM, SMART, PANTHER, PRINTS, TIGR, TIGRFAMS, HSSP, HAMAP, PRODOM, KEGG, KEGG compound, KEGG drug, PUBCHEM databases. The integration with the databases allows to obtain comprehensive information about the structural-functional-evolutionary features of proteins, their sites and also ligands.

Application of the PDBSite database and the PDBSiteScan tool to drug design. Let us use Leptin as an example of how we applied the program resources are developed for the search of small molecule compounds that might be candidates for drug design. Leptin is a protein of great interest in medical research (Peelman *et al*., 2004). We analysed the potential capacity of biologically active molecular compounds from the PDB database to bind to leptin. Leptin is an adipocyte derived hormone that circulates in the serum in the free and bound form. Serum levels of leptin reflect the amount of energy stored in adipose tissue. Short-term energy disbalance, as well as serum levels of several cytokines and hormones, influence circulating leptin levels. Leptin acts by binding to specific receptors in the hypothalamus to alter the expression of several neuropeptides that regulate neuroendocrine function and energy intake and expenditure. Thus, leptin plays an important role in the pathogenesis of obesity and food intake disorders and it is thought to mediate the neuroendocrine response to food deprivation.

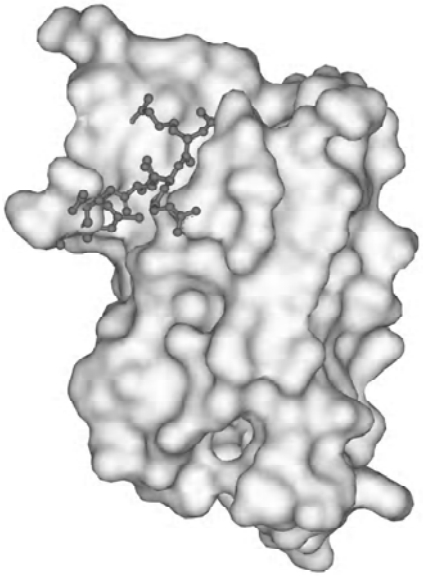


Figure 1. Potential leptin – ACE-ARG-ARG-LEU-ASN-FCL-NH peptide complex. Leptin is shown as surface molecule, the peptide is depicted, using ball and stick model.

It was found that the ACE-ARG-ARG-LEU-ASN-FCL-NH peptide, developed for the inhibition of the cyclin-dependent kinase 2/cyclin complex (Kontopidis *et al*., 2003) is also capable of binding to leptin (Fig. 1). Peelman (Peelman *et al*., 2004) have demonstrated that mutation at positions 41, 115–118, 122 and 124 of leptin affect its binding to the membrane proximal cytokine receptor homology domain (CRH2). CRH2 is the domain of the leptin receptor. The binding site of ACE-ARG-ARG-LEU-ASN-FCL-NH to leptin covers these positions. It can be thus suggested that the peptide can inhibit binding leptin to receptor. The current results may be a good start for further modification of the peptide with the aim of abolishing its binding capacity to cyclin and, moreover, to enhance its specific binding to leptin.

ACKNOWLEDGEMENTS

Work was supported in part by Russian Foundation for Basic Research No. 05-04-49283, the State contract No. 02.434.11.3004 of 01.04.2005 and No. 02.467.11.1005 of 30.09.2005 with the Federal Agency for science and innovation "Identification of promising targets for the action of new drugs on the basis of gene network reconstruction", Interdisciplinary integrative project for basic research of the SB RAS No. 115 "Development of intellectual and informational technologies for the generation and analysis of knowledge to support basic research in the area of natural science", innovation project of Federal Agency of Science and innovation IT-CP.5/001 "Development of software for computer modeling and design in postgenomic system biology (system biology *in silico*)" and the CRDF Rup2-2629-NO-04.

REFERENCES

Hendlich M. *et al*. (2003) Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *J. Mol. Biol.*, **326**, 607–620.

Jones S. *et al*. (2003) Using structural motif templates to identify proteins with DNA binding function. *Nucl. Acids Res*., **31**, 2811–2823.

Ivanisenko V.A. *et al*. (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. *Nucl. Acids Res.*, **32**, W549–W554.

Ivanisenko V.A. *et al*. (2005) PDBSite: a database of the 3D structure of protein functional sites. *Nucl. Acids Res.*, **33**, D183–D187.

Ko J. *et al*. (2005) Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics*, **21**, i258–i265.

Kontopidis G. *et al*. (2003) Insights into cyclin groove recognition: complex crystal structures and inhibitor design through ligand exchange. *Structure (Camb)*, **11**, 1537–1546.

Laskowski R.A. *et al*. (2005a) Protein function prediction using local 3D templates. *J. Mol. Biol.*, **351**, 614–626.

Laskowski R.A. *et al*. (2005b) ProFunc: a server for predicting protein function from 3D structure. *Nucl. Acids Res*, **33**, W89–W93.

Peelman F. *et al*. (2004) Mapping of the leptin binding sites and design of a leptin antagonist. *J. Biol. Chem.*, **279**, 41038–41046.

Shehadi I.A. *et al*. (2005) Active site prediction for comparative model structures with thematics. *J. Bioinform Comput. Biol.*, **3**, 127–143.

PROTEIN-PROTEIN INTERACTIONS AS NEW TARGETS FOR DRUG DESIGN: INTERACTIVE LINKS BETWEEN VIRTUAL AND EXPERIMENTAL APPROACHES

Ivanov A.S.*[*], *Gnedenko O.V., Molnar A.A., Mezentsev Yu.V., Lisitsa A.V., Archakov A.I.
V.N. Orechovich Institute of Biomedical Chemistry, RAMS, Moscow, 119121, Russia
[*] Corresponding author: e-mail: alexei.ivanov@ibmc.msk.ru, ivanov@ibmh.msk.su

Key words: protein-protein interaction, molecular modeling, computer-aided drug design, platform "from gene to drug", optical biosensor, SPR

SUMMARY

Motivation: Protein-protein and protein-ligands interactions play a central role in biochemical reactions, and understanding these processes is an important step in several fields of biomedical science and drug discovery.

Results: Our research is conducted on a number of protein-protein interactions.

We attempt in this report to show interactive links between virtual and experimental approaches in total pipeline "From gene to drug and using modern SPR technology (optical biosensor) for assessing the strengths of protein-protein and protein-ligand interactions.

Availability: Preprint of this paper is available on request from the authors.

INTRODUCTION

Genome sequencing has provided fast growth of our knowledge about proteins present in different live organisms. However this data tell us rather small information about the function of proteins because they often work in complex assemblies of several macromolecules and small ligands. Such complexes play crucial roles in most cellular processes and widely diverse in their activity and size.

Structural and functional description of protein-protein interactions is an important step toward understanding of biological processes. The applied area of such exploration is searching of new targets and creation new generation of highly effective and safety drugs.

Currently there are about 35 000 known structures in PDB, among them about 12 000 structures involving two or more protein chains. Within protein-protein complexes, two different types can be distinguished, homo- and hetero-complexes. It is known from PDB statistics that homo-complexes often exist as dimers, comparatively uncommon – as tetramers and very rare – as trimers or high multimeric complexes.

The contact surfaces of the protein complexes have unique structure and properties, so they represent prospective targets for a new generation of drugs (Veselovsky, 2002). Currently many investigations were undertaken to find or design small molecules that block protein-protein (protein-peptide) interactions (Pagliaro *et al*., 2004) and in particular protein dimerization (oligomerization). We were intrigue to investigate the mechanism of protein-protein interactions and to apply the gained knowledge towards drug design. Our research is conducted on a number of protein-protein interactions. We attempt in this report to show interactive links between virtual and experimental approaches in total pipeline "From gene to

drug" (Fig. 1) (Veselovsky, 2003; Ivanov, 2005) and using modern SPR technology (optical biosensor) for assessing the strengths of protein-protein and protein-ligand interactions.

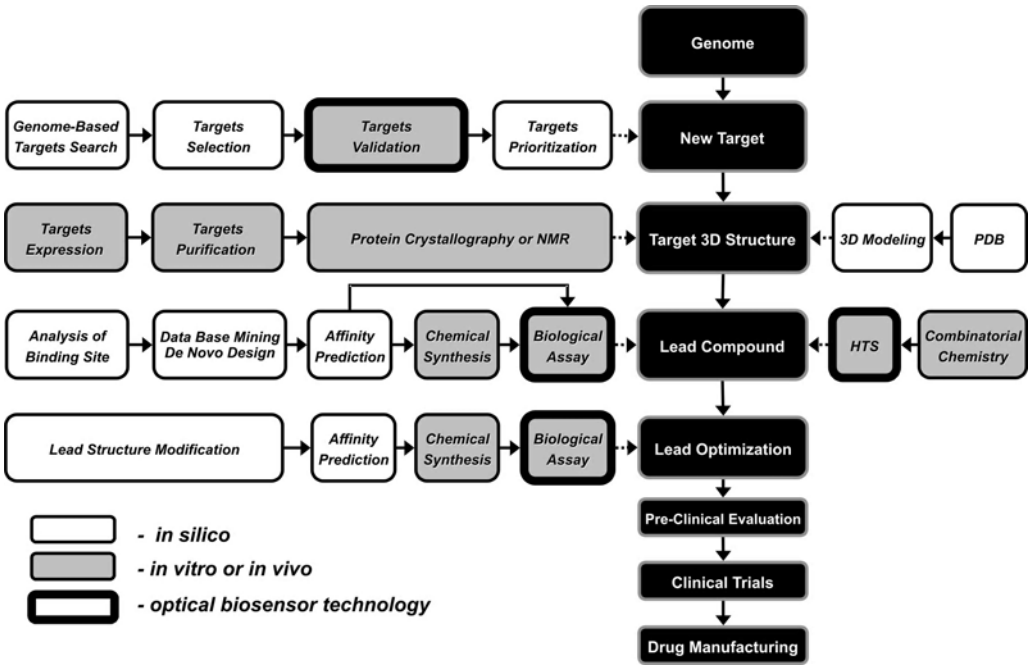


Figure 1. Pipeline "From gene to drug": integration of virtual and real experiments.

METHODOLOGY

Bioinformatics* (*in silico*) *approaches. Bioinformatics methods and molecular modeling software provide useful tools to help researchers elucidate protein interaction mechanisms by generating 3D models of intermolecular complexes and using scoring functions to select the most likely molecular complex hypothesis and discovering of lead candidates as inhibitors of protein-protein interaction.

Experimental (in vitro) approaches. Several proteomics technologies have been developed and adapted to investigate protein-protein interactions. The yeast two-hybrid method allows the mapping of binary or pair-wise interactions, protein chips are suited to detect protein-protein, protein-lipid and protein-ligand interactions. Affinity capturing method based on the chip of optical biosensor (fishing) was coupled to mass spectrometry (MS) protein identification techniques for identification of partners in bimolecular or multimolecular protein complexes. Here, we will highlight the universal character of optical biosensor based on surface plasmon resonance technology (SPR) for solving different experimental tasks in analysis of protein-protein and protein-ligand interactions (McDonnell, 2001).

IMPLEMENTATION

Analysis of oligomerization of L-asparaginase. This enzyme is widely used in medical practice as therapeutic agents for treatment acute leukemia. However its application is accompanied by several side effects that caused by insufficient enzyme selectivity. The last one is defined by structure of the active site located between subunits of protein tetramer. Hence, the process of asparaginase oligomerization plays a key role in

formation of the active site and defines substrate specificity. In the present work we modeled spatial structure of L-asparaginase from *Erwinia carotovora* (Fig. 2) based on homology with L-asparaginase from *Erwinia chrysanthemi* and the comparative analysis of the interface between subunits was done.

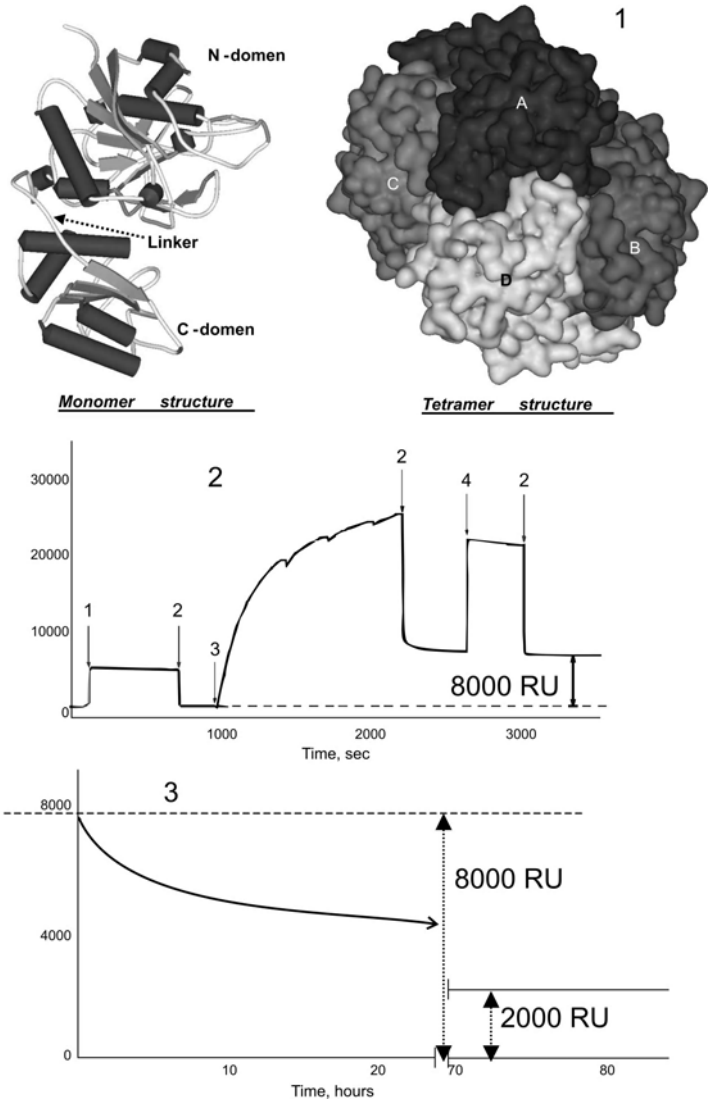


Figure 2. Experiments with L-asparaginase. (1) – 3D models of monomer and tetramer; (2) – tetramers immobilization on CM5 chip; (3) – sensogram of tetramers dissociation up to monomer.

We also developed experimental approach to study the process of oligomerization of this enzyme using optical biosensor Biacore 3000. Protein was immobilized on a surface of optical chip CM5 and tetramers dissociation up to monomeric condition has been registered. The subsequent restoration of enzyme tetramers was also carried out.

HIV-1 protease (HIVp) dimerization. The main function of HIVp is the slicing viral preprotein on mature proteins. The enzyme also aggravates AIDS by damaging the host cell proteins. Many rather effective competitive inhibitors of HIVp are known and some of them are used now in AIDS therapy. Their systematic application as the drugs, however, inevitably promotes the generation of the viral strains that are resistant both to the inhibitor used and to most of its structural analogs. The drug-resistant protease modification is a result of the point mutation i.e. replacement of one amino acid residue in

both identical enzyme subunits. HIVp operates in homodimeric form, each of identical subunits being consisted of 99 amino acid residues. The main interface region in the homodimer represents the antiparallel four-strand β-sheet, which involves the C- and N-terminal peptides of both subunits (Fig. 3).

It is natural to assume that some ligand binding with any subunit can interfere with subunit dimerization. If the binding site coincides or overlaps with the interface region, all the mutations that diminish subunit affinity to a ligand will be also affect negatively inter-subunit interactions. As a result the mutant protease will be form less stable and, consequently, less active dimers. At least two highly specialized and synchronous mutations are necessary to obtain a drug-resistant strain with high inter-subunit affinity. It is obvious that such mutations are highly improbable.

Some years ago we have begun the project on designing inhibitors of HIVp dimerization that are not capable to stimulate the appearance of drug-resistant viral strains. There are few general strategies for the generation of synthetic molecules that directly modulate protein-protein interactions. We have implemented *de-novo* design using molecular modeling software Sybyl (Tripos Inc.). Constructed structures of lead compounds (peptidomimetic inhibitors of HIVp dimerization) currently are under synthesis. It was necessary to develop biological assay for direct *in vitro* analysis of interactions of lead compounds with interface site of HIVp monomer. This assay was created based on optical biosensor Biacore 3000. HIVp was immobilized in dimeric form in two channels on optical chip CM5. Than protein dimers in channel 1 were stabilized by chemical cross-linking, while in channel 2 HIVp dimers were dissociated up to monomers. Assay trial experiments were carried out with known test peptide inhibitor (Fig. 4).

It is visible, that inhibitor interacts only with monomeric form of HIVp, which indicate that molecules of inhibitor bind only to subunits interface.

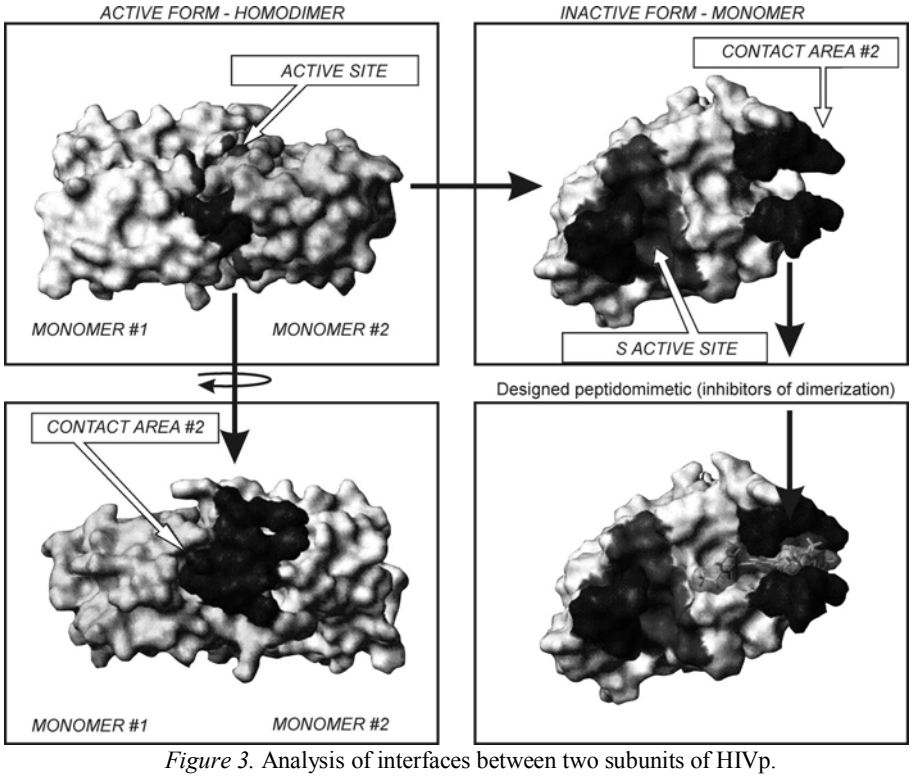


Figure 3. Analysis of interfaces between two subunits of HIVp.

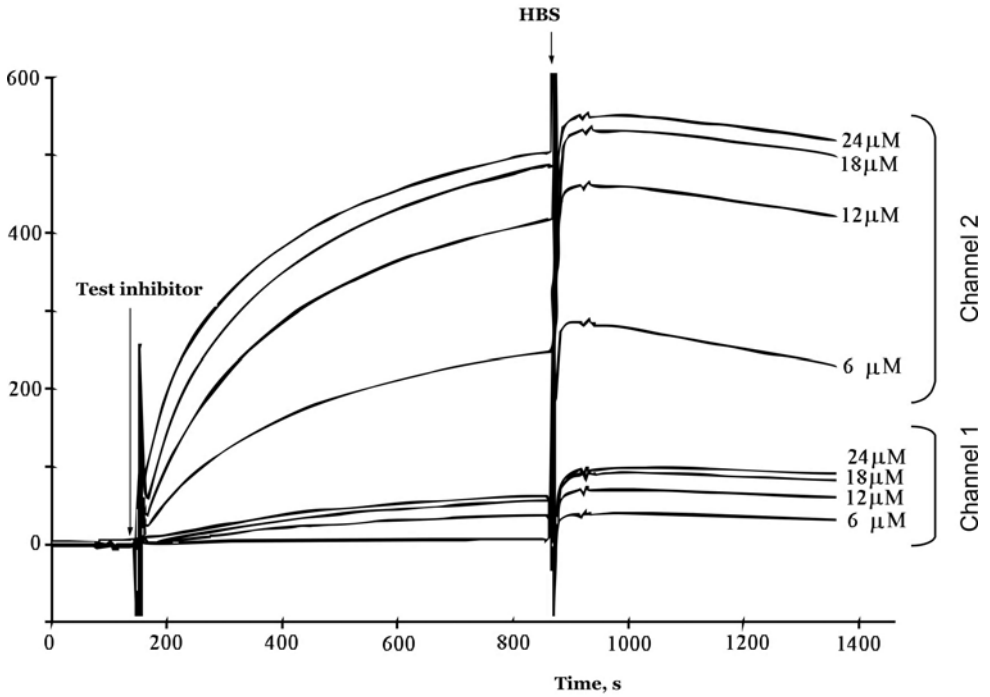


Figure 4. In vitro assay for inhibitors of HIVp dimerization.

REFERENCES

Ivanov A.S. *et al.* (2005) Bioinformatics platform development: from gene to lead compound. In Larson R.S. (ed.), *Methods in Molecular Biology, vol. 316: Bioinformatics and Drug Discovery*. Humana Press, Totowa, pp. 389–432.

McDonnell J.M. (2001) Surface plasmon resonance: towards an understanding of the mechanisms of biological molecular recognition. *Cur. Op. Chem. Biol.*, **5**, 572–577.

Pagliaro L. *et al*. (2004) Emerging classes of protein–protein interaction inhibitors and new tools for their development. *Cur. Op. Chem. Biol*., **8**, 442–449.

Veselovsky A.V. *et al.* (2002) Protein-protein interactions: mechanisms and modification by drugs. *J. Mol. Recognit.*, **15**, 405–422.

Veselovsky A.V., Ivanov A.S. (2003) Strategy of computer-aided drug design. *Current Drug Targets – Infectious Disorders*, **3**, 33–40.

THE CONTRIBUTION OF ALTERNATIVE TRANSLATION START SITES TO HUMAN PROTEIN DIVERSITY

Kochetov A.V. [*1], ***Sarai A.***[2], ***Kolchanov N.A.***[1]
[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia; [2] Kyushu Institute of Technology, Dept. Biochemical Engineering and Science, Iizuka, 820-8502 Japan
[*] Corresponding author: ak@bionet.nsc.ru

Key words: mRNA, alternative translation, proteome, organelle

SUMMARY

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

Results: Sequence organization of translation initiation signal of human mRNAs was analyzed. It was found that a suboptimal context of annotated start codon correlated with a significantly higher frequency of in-frame downstream AUG codons. We compared predicted subcellular localizations of annotated human proteins and their potential N-terminally truncated forms started from the nearest downstream in-frame AUG codons. It was found that the localization of full and N-truncated protein variants was often different: ca. 3.5 % of human genes tested could produce additional proteins with other targeting signals. It is likely that the in-frame downstream AUGs may be frequently utilized to synthesize additional proteins possessing new functional properties and such a translational polymorphism may serve as an important source of cellular and organelle proteomes.

INTRODUCTION

Human genome was assumed to contain ca. 20,000–25,000 protein coding genes. The number of proteins actually formed may be considerably higher because of alternative splicing. Another possible source of new protein forms is translational heterogeneity where several AUG codons within mRNAs may serve as alternative translation start sites (TSSs) to produce overlapping proteins displaying different properties (e.g., Bab *et al*., 1999; Watanabe *et al*., 2001). The contribution of such a translational polymorphism to proteome complexity was not evaluated.

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

It was found that a part of cellular mRNAs with start AUG codon lying in suboptimal context is relatively large as well as a part of mRNAs with AUG-containing 5′ untranslated regions (5′-UTRs) (Rogozin *et al*., 2001). It is likely that at least some mRNAs with suboptimal start codon context may produce two or more functional polypeptides. To test this assumption we isolated a sample of human cDNAs and

compared predicted subcellular localizations of polypeptides started from either annotated suboptimal TSS or the nearest downstream in-frame AUG codons.

METHODS

32451 GenBank entries were obtained at http://www.ncbi.nlm.nih.gov/ using the following search fields: "*Homo sapiens* AND complete CDS"; Limits: "mRNA; Genomic DNA/RNA, excluding ESTs, STSs, GSS, working draft, and patents". Of them, 27616 sequences contained both the complete coding parts and 5′ UTRs shorter than 1000 nucleotides in length. Subcellular localizations of proteins were evaluated by TargetP prediction progam (Emanuelsson *et al*., 2000) used with default parameters.

RESULTS AND DISCUSSION

Analysis of nucleotide frequencies in AUG context positions showed that 17 % of human mRNAs contained annotated start codon in a suboptimal context (*i.e.*, they contained pyrimidines in position –3) and 44 % of human mRNAs contained AUG(s) within annotated 5′-UTRs. Despite a high uAUG content, the observed average uAUG frequency (8 AUGs per 1000 nucleotides) was lower than the expected value (12 AUGs per 1000 nucleotides; calculated as a product of frequencies of A, T, and G), which may reflect the selection against the presence of AUGs within 5′ UTR of eukaryotic mRNAs (Rogozin *et al*., 2001).

We calculated the average frequencies of in-frame AUG codons at the CDS beginning (from 3rd to 9th codons) downstream of annotated start site. It was found that average AUG frequency downstream the TSS in optimal context (purine in pos. -3) was significantly lower than in mRNAs with a suboptimal start codon context (pyrimidine in pos. -3): 0.016 *versus* 0.025, respectively. The difference was not observed for downstream AUG triplets located out of the CDS frame (0.018 *versus* 0.017). This may mean that the in-frame AUGs located downstream of a "weak" TSS may be of functional importance. To test this assumption we prepared two samples of proteins: started either from annotated TSS in a suboptimal context or from the nearest downstream in-frame AUG codons. In total, subcellular localization of 3327 full and N-truncated proteins were compared.

The results of prediction are shown in Table 1. One can see that N-truncated forms of many secreted polypeptides (18 %) lose their targets. It was expected since N-truncated polypeptides could lose their secretory leader peptides. However, 10 % of N-truncated proteins acquire sorting signals *de novo* (predicted localization was changed from "Other" to "MTP" or "SP") and 2 % change their predicted subcellular locations (from mitochodria to secretory pathway or *vice versa*; detailed description is available by request). It may mean that a substantial part of human mRNAs produce two (or more?) proteins each with a specific subcellular localization due to alternative translation.

Alternative translation allows to generate two or more protein forms. It might represent an appropriate way to address proteins of the same function to different locations or generate protein forms with different functions (e.g., Bab *et al*., 1999; Watanabe *et al*., 2001). Note, that the ratio of full- and N-truncated protein variants may be tightly regulated through adjustment of the start codons contexts to control initiation/scanthrough ratio (Kozak, 2002) that may provide a unique mechanism of expression control. It was found that 30 % of the N-truncated proteins were targeted to other cellular compartments. Thus, *ca*. 12 % of human cDNA analyzed might encode N-truncated protein variants and *ca*. 3.5 % might encode additional protein variants targeted to other subcellular compartments.

According to recent evaluations, human genome contains as many as 20,000–25,000 genes (International human genome sequencing consortium, 2004). Of them, about 3000 genes might encode additional protein variants due to the translational polymorphism. Such a mechanism could make an important contribution to human cellular and organelle proteomes.

Table 1. Subcellular localization of human annotated proteins and their putative N-truncated variants (%) predicted with *TargetP* program (Emanuelsson *et al*., 2000)*

Annotated		N-truncated		
Location	Size of fraction	MTP	SP	Others
MTP	13	3	1	9
SP	17	1	7	9
Other	70	5	5	60
Total	100	9	13	78

* MTP, mitochondria targeting peptide; SP, secretory peptide.

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

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (05-04-48207) and RAS program (Dynamics of Plant, Animal and Human Gene Pools). We thank SD RAS Complex Integration Program (N5.3), and Ministry of Industry, Sciences and Technologies of Russian Federation (2275.2003.4) for partial support.

REFERENCES

Bab I., Smith E., Gavish H., Attar-Namdar M., Chorev M., Chen Y.C., Muhlrad A., Birnbaum M.J., Stein G., Frenkel B. (1999) Biosynthesis of osteogenic growth peptide via alternative translational initiation at AUG85 of histone H4 mRNA. *J. Biol. Chem.*, **274**, 14474–14481.

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

International human genome sequencing consortium (2004) Finishing the euchromatic sequence of the human genome. *Nature*, **431**, 931–945.

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

Kozak M. (2002) Pushing the limits of the scanning mechanism for initiation of translation. *Gene*, **299**, 1–34.

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

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

STRUCTURAL DETERMINANTS OF CARDIOTOXINS MEMBRANE BINDING: A MOLECULAR MODELING APPROACH

Konshina A.G.*, *, Dubinnyi M.A., Efremov R.G.
Shemyakin–Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russia
* Corresponding author: e-mail: nastya@nmr.ru

Key words: membrane-protein interaction, implicit model of membrane, cytotoxin, molecular dynamics

SUMMARY

Motivations: Because of experimental difficulties with characterization of membrane – protein interactions, development of molecular modeling approaches is a field of especial interest. To understand how the membrane binding occurs on molecular level and what structural features are responsible for the strength of binding, we applied a new algorithm, combining molecular dynamics (MD) simulations in water followed by Monte-Carlo (MC) search in implicit membrane, to particular biological objects – two homologous cardiotoxins (CTs), CTI and CTII, from snake venom.

Results: MD simulations of quite structurally rigid molecules of toxins show that CTs differ significantly in structural stability of their loops I and II.

In order to assess the mode of membrane binding for different CTs, their NMR- and MD-derived models were futher employed in MC search with implicit membrane. The results obtained reveal the exclusive role of a charged residue in loop II and minor local differences in structure of toxins in their mode of membrane binding.

It is proposed that a long-living water molecule found in the loop II of CT I may play a role in regulating the lipid binding mode of CTs.

INTRODUCTION

A large number of protein molecules are produced by cell either for own cytoplasmic membrane or for interaction with other ones. From the latest – a fair amount of toxins and all of them should have certain adaptive features for efficient overcoming of membrane barrier to damage the cell. One of the interesting toxins' groups is β-structured cardiotoxins from snake venom. CTs belong to the family of the "three-finger" proteins whose fold consists of three β-stranded "finger-shaped" loops protruding from a globular core with four disulfide bridges. Despite the similarity of their spatial models and high level of sequence homology, CTs are characterized by a variety of biological activities and essentially differ in cytotoxicity (Kumar *et al*., 1997)). It is established now that CTs have unspecific cytolytic effect but little is known about the mode of toxins action on membranes.

Structural and evolutionary comparisons among the CTs family indicated that the major structural plasticity accompanied with the hypervariable amino acid composition is present at the tips of the loops I and II. Also, in a number of three-dimensional (3D) models of CTs a long-living water molecule was found to be tightly hydrogen bonded to the residues of loop II (Sue *et al*., 2001). This loop has been proposed to be the most important cytolytic domain. CTs have been divided into P- and S-types, depending on the

presence of either Pro31 or Ser29 residues at the tip of loop II. It was found that the P-type CTs interact with bilayers stronger than those of the S-type (Chien *et al*., 1994; Dubovskii *et al*., 2005).

Based on NMR data in solution and DPC micelles, spatial structures of both types of CTs, isolated from the same cobra (*Naja oxiana*), CT I (S-type) and CT II (P-type), were established in our laboratory. It was found that they are similar. Moreover, the bound water molecules were identified in their loop II. At the same time, toxins show different degree of cytolytic activity (Feofanov *et al*., 2004).

It is suggested that the subtle conformational differences in loops extremities as well as possible participation of the bound water may define effectiveness of membrane binding of CTs. Difficulties in getting of such delicate structural information via experiments, make the molecular modeling approaches especially actual. To understand the structural features that may promote the differences in mode of membrane binding, we performed MD and MC simulations of both toxins, respectively in water and in implicit membrane.

METHODS

MD in water was used to explore conformational possibilities of CTs as well as to find reliable starting structures for subsequent modeling of CTs binding to membrane using MC conformational search in water-membrane environment.

The spatial models of CT I and CT II determined by NMR spectroscopy in aqueous solution were used as starting structures. Three MD trajectories (~10–22 ns) were obtained for each toxin. In each case stepwise energy minimization and linear heating of the system were preceding the collection run. MD calculations and data processing were performed using the GROMACS v3.1.4 software and a set of original programs. Several MD-conformers as well as experimentally determined 3D models of CTs were used as starting conformations in MC simulations with implicit three-layer membrane model (water–cyclohexane–water). Previously, we have designed a model of the implicit membrane in which the solvent effect was established by the addition of a special term based on the use of empirical atomic solvation parameters incorporated into the potential energy function of a protein in "vacuum" (Efremov *et al*., 2004). The starting conformations were arbitrarily placed in water, and several successive MC calculations (3–5 $x10^3$ steps each) were carried out with sampling of 1–2 randomly chosen dihedral angles. Resulting low-energy states in the range of 10 kcal/mol from the minimal energy state were analyzed. Calculations were performed using the FANMEM program (a modified version of the FANTOM package).

RESULTS

Analysis of MD data revealed differences in the conformational lability of toxins loop I-II regions. The structure of the loop I of CT II was found to be more stable than that of CT I. On the contrary, dynamic features of CT II may be mainly characterized in terms of structural flexibility of the loop II: some families of MD-conformers differed markedly by conformation of this protein part.

A single water molecule was found to be tightly hydrogen bonded to the following sites of CT I: M26:NH, D29:O (or OD1, OD2), and I32:O. A number of such waters with long residence time (more than nanosecond) were observed in MD simulations. The water molecules bind preferentially in this site, and very rare – in other positions of loop II. To the contrary, multiple binding sites for waters were detected in CT II. Moreover, in a large majority of accumulated MD conformers no bound waters were found in this site.

Analysis of MC data shows that the geometry and the depth of CTs' insertion are determined by the location of hydrophobic loop's residues (which form apolar surface like a "bottom") relative to the positively charged conservative residues flanking the ends of loops I-III. Regardless of the starting structures, CT II inserts into bilayer with the hydrophobic extremities of its loops I-III. In case of CT I the mode of binding (via one, two or all three loops) is defined by the conformation of loop II. More precisely, this is determined by the location of the side chain of D29 with respect to the hydrophobic stretch formed by the loop's residues. In the low energy states this residue is always placed either on the membrane interface or in water.

DISCUSSION

Recent studies have shed some light on the structure-activity relationships of CTs. Thus, it has been revealed that the hydrophobic tips of loops I-III represent an important functional motif for binding of CTs to lipid bilayers. Indeed, the results of MC simulations have demonstrated that for efficient penetration into the membrane, the molecule must be able to form a continuous "hydrophobic bottom". It was proposed that the membrane binding is correlated with the ability of loop II to adopt a Ω-shaped conformation. This promotes formation of a single hydrophobic path ("bottom") by the loops I-III. Thus, for strong binding with membrane the loop II of CTs must contain mainly apolar residues. Also, it should have some features constraining its conformational mobility to favor the "right" conformation of this loop on the membrane interface. Indeed, the loop II of CT II is quite hydrophobic and has residue P30 (like other representatives of the P-type CTs) at its tip. Finding the bound waters among several P-type CTs suggests that this proline residue should play an important role in formation of the water binding Ω-shaped loop II. For the P-type toxin, CT II, effectiveness of its membrane binding was confirmed by a series of MC searches starting from several distinct conformational states founded in MD.

As shown from MC data analysis, the presence of charged residue D29 in the loop II of CT I substantially confines a number of states that may realize the hydrophobic stretch. As a consequence, the mode of membrane binding via one or two loops appeared among the low-energy states of CT I.

Recent NMR studies of CTI revealed the presence of a bound water molecule located near the tip of loop II and its absence in aqueous and micellar environment, respectively. The results of MD simulations of CT1 are completely consistent with the experimentally derived information. Moreover, the water binding site identified with the two independent techniques is identical.

It seems that the absence of Pro30 can make this loop too flexible. But the additional hydrogen bonds holding water molecules near the residues M26, D29, and I32 make the structure of loop II more rigid, thus avoiding significant conformational changes. Indeed, as seen from MD data, the loop II of CT I was less mobile as compared with that of CT II, where bound water molecules were observed much rarely. Comparison of the two NMR-derived 3D models of CT I, in solution and in membrane-like environment (DPC micelles), reveals only one substantial difference between the structures – namely, the conformation of the tip of loop II and, as consequence, opposite orientations of side chains of D29. Note, that the NMR-structure in DPC micelle does not contain a water binding site. The geometry of binding of such NMR-structure was similar to that of CT II: all three loops interact with the bilayer.

MC simulations with implicit membrane suggest the role of this particular conformation of loop II. Probably, it provides favorable orientation of the charged group of D29: away from the "hydrophobic bottom" formed by the extremities of loops I-III. We hypothesize that preserving of the definite loop II conformation through the binding

of water molecule favors its conformational "switching" into the "right" conformation (accompanied with the release of water) on the water-membrane interface.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grants 04-04-48875-a, 05-04-49283-a) and by the Russian Federation Federal Agency for Science and Innovations (The State contract 02.467.11.3003 of 20.04.2005, grant SS-4728.2006.4).

REFERENCES

Chien K.Y. *et al.* (1994) Two distinct types of cardiotoxin as revealed by the structure and activity relationship of their interaction with zwitterionic phospholipid dispersions. *J. Biol. Chem*., **269**, 14473–14483.

Dubovskii P.V. *et al.* (2005) Interaction of three-finger toxins with phospholipid membranes: comparison of S- and P-type cytotoxins. *Biochem J.*, **386**, 1–9.

Efremov R.G. *et al.* (2004) Peptides and Proteins in Membranes:What Can We Learn *via* Computer Simulations. *Current Med. Chemistry*, **11,** 2421–2442.

Feofanov A.V. (2004) Comparative study of structure and activity of cytotoxins from venom of the cobras *Naja oxiana*, *Naja kaouthia*, and *Naja haje*. *Biochemistry (Mosc)*, **69**, 1148–1157.

Kumar T. *et al.* (1997) Snake venom cardiotoxins-structure, dynamics, function and folding. *J. Biomol. Struct. and Dyn*., **15**, 431–463.

Sue S.C. *et al.* (2001) Dynamic characterization of the water binding loop in the P-type cardiotoxin: implication for the role of the bound water molecule. *Biochemistry*, **40**, 12782–12794.

OPTIMIZATION OF ACCURACY AND CONFIDENCE FOR ALIGNMENT ALGORITHMS EXPLOITING DATA ON SECONDARY STRUCTURE

***Litvinov I.I.*[*1, 3], *Finkelshtein A.V.*[2], *Roytberg M.A.*[*1, 3]**
[1] Institute of Mathematical Problems in Biology, RAS, Pushchino, Moscow Region, 142290, Russia;
[2] Institute of Protein Research, RAS, Pushchino, Moscow Region, 142290, Russia; [3] Pushchino State University, Pushchino, Moscow Region, 142290, Russia
[*] Corresponding authors: e-mail: mroytberg@mail.ru

Key words: protein sequence alignment, secondary structure prediction, alignment quality, accuracy, confidence

SUMMARY

Motivation: The quality of protein sequences alignment is a similarity between the alignment and the "golden standard" alignment reflecting the evolutionary history. The quality of algorithmically obtained alignment is crucial for many bioinformatics tasks.

Two main measures of alignment quality are *accuracy*, i.e. part of correctly restored positions of the golden standard alignment, and *confidence*, i.e. part of positions of the algorithmic alignments that belong to the golden standard alignment. The measures often are contradictory, i.e. the parameters optimizing one of the measures can result in low values of another.

Results: We have performed detailed investigation of accuracy and confidence of alignments obtained by different methods with different values of parameters. It was shown that the methods exploiting information about the secondary structure admit the simultaneous optimization of alignment accuracy and confidence with the same parameters values. This contrasts with the behavior of alignment accuracy/confidence for classic Smith-Waterman method.

INTRODUCTION

Pair-wise alignment of amino acid sequences is a core of many bioinformatics methods. The ideal goal of all alignment algorithms is to find a biologically correct alignment reflecting the evolutionary history of homologous proteins (Sunyaev *et al*., 2004); i.e. aligned positions have to correspond to the same position of their common ancestor. The "quality" of an algorithmic alignment of amino acid sequences (i.e., its similarity to the biologically correct alignment) is critical for many applications, e.g. homology modeling, database homology search, protein domains analysis, etc. Biologically correct alignment is unknown, thus to measure the alignment quality one has to use an approximation of the biologically correct alignment as the "golden standard". Since the tertiary structure of proteins is much more conservative than their sequences, we use the alignments obtained by superimposing the protein spatial structures as a "golden standard". Alignment quality can be described by two complementary measures: *accuracy* (a number of identically aligned positions in algorithmic and reference alignment divided by total number of positions aligned in algorithmic alignment) and *confidence* (a number of identically aligned positions in the algorithmic and reference alignment divided by total number of positions aligned in the reference alignment).

The quality of algorithmic alignments crucially depends on the similarity of the sequences to be compared. For instance, the accuracy of the Smith-Waterman (SW) algorithm is 84 % when the protein identity (i.e., the portion of identical positions in two proteins) is no less than 30 %; and if the identity is below 30 %, the alignment accuracy is about 30 % (Sunyaev *et al*., 2004). The rapid approximate alignment algorithms, such as BLAST and FASTA, are even less accurate. To improve the accuracy of algorithmic alignments one can use combined methods taking into account both sequences and the (predicted) secondary structures. E.g. we have proposed the method STRUSWER (Litvinov *et al*., 2006) algorithm, which utilizes an additional bonus for matching identical elements of secondary structures; secondary structures can be determined experimentally or theoretically. Another method of this type is the Wallqvist-Fukunishi-Murphy-Fadel-Levy algorithm (WFMFL) (Wallqvist *et al*., 2000).

The optimal values of parameters depending on the protein sequence identity were found for all above algorithms. However, the two measures of alignment quality usually lead to different values of parameters. E.g. it is common knowledge that the alignment confidence is more essential for the database search, but the parameter values optimizing the confidence results in very low values of the accuracy.

The aim of the presented work was detailed investigation of the dependence of alignment quality on the algorithm parameters. We show that unlike the classic alignment algorithms (Smith-Waterman, etc) the secondary structure based methods allow simultaneous optimization of accuracy and confidence.

MATERIALS AND METHODS

Secondary structure. To predict the secondary structure we have used the PSIPRED program (Jones, 1999). The data presented below were obtained with the full version (prediction based on preliminary homology search) and the deterministic representation of the prediction (each residue is assigned with one of three letters: H (helix), E (beta) and L (loop). The other modes of the PSIPRED program as well as usage of experimentally obtained secondary structures from the DSSP database lead to the similar results.

Golden standard alignments. As a golden standard, we used manually verified structure alignments from the BAliBase (Bahr *et al*., 2001) protein structure database, as a source of "golden standard" alignments. We have used alignments from BAliBase Reference 1, the sequence identity level for the Reference is mainly 10–50 %. The test set was consisted of all protein pairs meeting following condition: both proteins belong to the same multiple alignment of BAliBase's Reference 1 and their 3D-structures are known.

Evaluation of the alignment quality. To compare two alignments (algorithmic and golden standard ones) and to estimate the agreement between them, we used two measures, accuracy and confidence. The alignment accuracy (Acc) was defined as a ratio of the number of positions (I) aligned similarly in the reference and algorithmic alignments to the number of aligned positions in the reference alignment (G): Acc = I/G.

The alignment confidence (Conf) was defined as a ratio of the number of positions aligned similarly in the reference and algorithmic alignments to the number of aligned positions in the algorithmic alignment (A): Conf = I/A.

Alignment algorithms utilizing the secondary structure data. We have tested two such algorithms, our algorithm STRUSWER (Litvinov *et al*., 2006) and the algorithm of Wallqvist-Fukunishi-Murphy-Fadel-Levy (WFMFL) (Wallqvist *et al*., 2000). STRUSWER is a modification of the Smith-Waterman (SW) algorithm. The only difference is that the score of the matching of i-th amino acid residue of one sequence with the jth residue of the other involves an extra summand SBON*SS[i, j], where SBON is a parameter of the algorithm and SS[i, j] = 1 if the residues are assigned with the same secondary structure type and the type is H or E; otherwise and SS[i, j] = 0. SW algorithm corresponds to SBON = 0.

The WFMFL algorithm modifies the Smith-Waterman algorithm in a similar way, but the extra summand is determined by the predefined 3 × 3 matrix depending on secondary structure types of compared residue (Wallqvist *et al*., 2000).

Optimization of the parameters of the program. Three algorithms were run for each pair of proteins from BAliBASE set and for each set of parameters: (1) SW algorithm (secondary structure disregarded, i.e. SBON = 0); (2) STRUSWER algorithm with the secondary structure predicted using the PSIPRED program; (3) WFMFL alignment with the secondary structure predicted using the PSIPRED program. Each algorithm was implemented with different values of parameters; the following integer values of parameters were checked: Gap Opening Penalty (GOP): from 4 to 20, Gap elongation penalty (GEP) from 1 to 7; SBON: from 1 to 30, GOP from 4 to 20, and GEP from 1 to 7. Thus, for each protein pair we have constructed 17 × 7 of SW and WFMFL alignments and 30 × 17 × 7 STRUSWER alignments (parameter SBON is applicable only for STRUSWER). Each of the algorithmic alignments was compared with the corresponding golden standard alignment, to obtain its accuracy and confidence. Finally, the results obtained for all protein pairs were averaged to yield average values <Acc> = <I/G> and <Conf> = <I/A> for a given algorithm and a set of parameters.

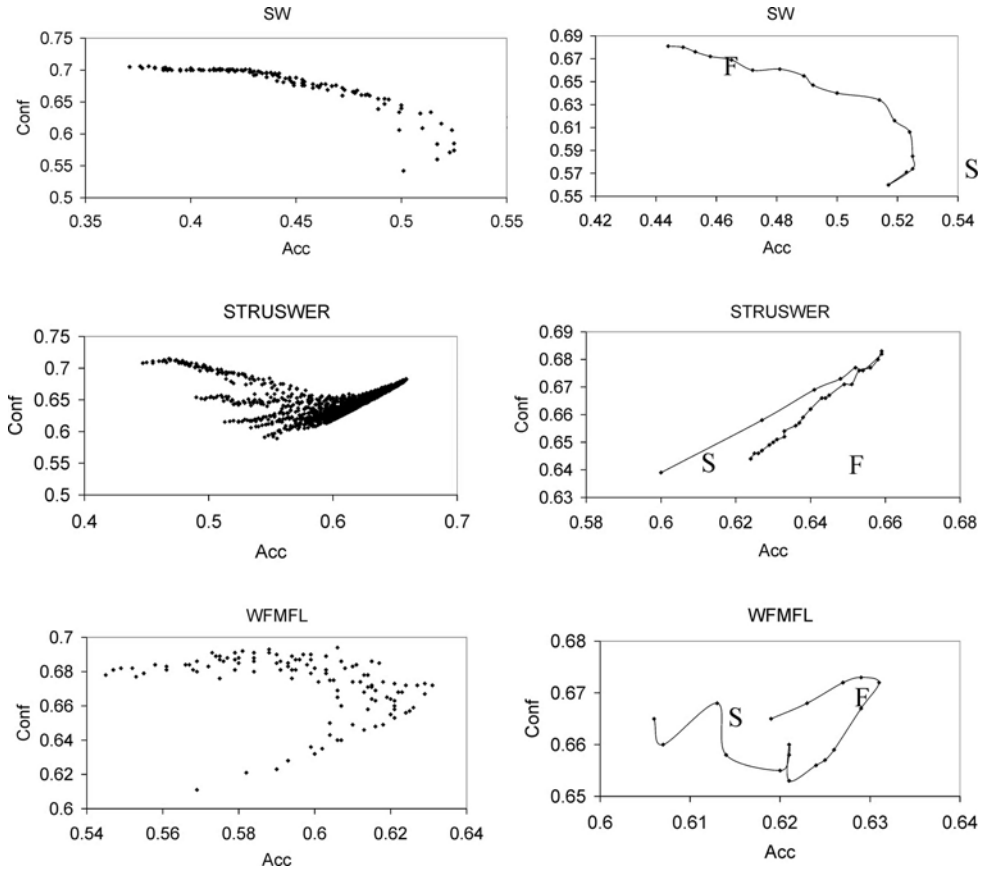


Figure 1. Full dataset. Accuracy/Confidence scatter-plots (left) and "trajectory" plots (right) for each method. Each point of the scatter-plot corresponds to a set of parameters GOP, GEP and SBON (the last for STRUSWER only). Each point on "trajectory" plot corresponds to a value of main parameter (SBON for STRUSWER; GOP for SW and WFMFL). Other parameters are chosen to optimize the average accuracy. Start (the lowest value of the main parameter) and finish (the greatest value) are marked with "S" and "F" respectively. X-axes presents the average accuracy and Y-axes presents the average confidence for a parameter set (see Materials and Methods).

RESULTS AND DISCUSSION

In our previous work (Litvinov *et al*., 2006) we have shown that methods using information about secondary structure provide essentially more accurate alignments than the Smith-Waterman algorithm. This advantage is the more valuable the lower is the sequence identity (see Fig. 2). Here we present the detailed investigation of the dependence of accuracy and confidence of alignments on the parameters of alignment algorithms. The most striking result is possibility to achieve both maximal accuracy and almost maximal confidence for the same values of STRUSWER parameters (see Fig. 1). The maximal value of confidence (0.7) corresponds to "strong" values GOP = 17; GEP = 6, SBON = 1 but it provides very low accuracy (0.47). Fortunately, "weak" parameters providing maximal value of accuracy (SBON = 8; GOP = 9; GEP = 1) correspond to the almost maximal value of the confidence (0.683 compared to 0.707). Fig. 1 (middle-right) shows how the accuracy depends on the SBON parameter. The method WFMFL (Wallqvist *et al*., 2000) demonstrate the similar behavior related to parameters GOP/GEP (see Fig. 1, bottom).

The optimal values of parameters (leading to acc= 0.63; conf = 0.67) are those maximizing accuracy and they coincide with the values recommended in (Wallqvist *et al*., 2000). In contrast the Smith-Waterman method that has only two parameters, does not allow simultaneous optimization of accuracy and confidence (see Fig. 1, top). The behavior of the algorithms' accuracy/confidence is essentially the same if we restrict ourselves with low-homology protein pairs (see Fig. 2). The optimal parameter values are almost the same.

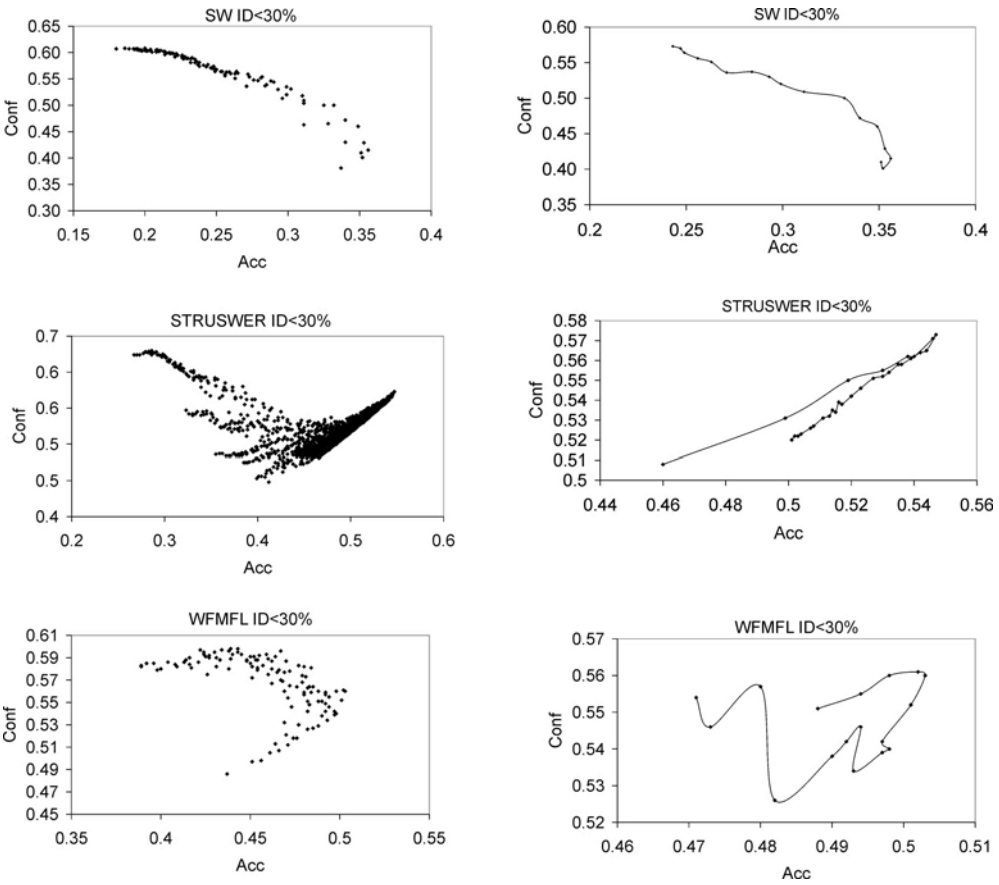


Figure 2. Accuracy/Confidence plots (see Fig. 1) for proteins with sequence similarity less than 30 %.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (project Nos 03-04-49469, 02-07-90412), by grant from the RF Ministry for Industry, Science, and Technology (20/2002, 5/2003), NWO, ECO-NET, and by the program of RF Ministry of Science and Education (contract No. 02.434.11.1008).

REFERENCES

Bahr A., Thompson J.D., Thierry J.-C., Poch O. (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. *Nucl. Acids Res.*, **29**, 323–326.

Jones D. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.

Litvinov I.I., Lobanov M. Yu., Mironov A.A., Finkelstein A.V., Roytberg M.A. (2006) Information about secondary structure improves quality of protein alignment. *Mol. Biol*. In press.

Sunyaev S.R., Bogopolsky G.A., Oleynikova N.V., Vlasov P.K., Finkelstein A.V., Roytberg M.A. (2004) From analysis of protein structural alignments toward a novel approach to align protein sequences. *Proteins*, **54**, 569–582.

Wallqvist A., Fukunishi Y., Murphy L.R., Fadel A., Levy R.M. (2000) Iterative sequence/secondary structure search for protein homologs: comparison with amino acid sequence alignments and application to fold recognition in genome databases. *Bioinformatics*, **16**, 988–1002.

DEVELOPMENT OF A HIERARCHICAL CLASSIFICATION OF THE TIM-BARREL TYPE GLYCOSIDE HYDROLASES

Naumoff D.G.
State Institute for Genetics and Selection of Industrial Microorganisms, Moscow, Russia,
e-mail: daniil_naumoff@yahoo.com

Key words: hierarchical classification, $(\beta/\alpha)_8$-barrel fold, enzyme classification, protein family, protein phylogeny, glycoside hydrolase, α-galactosidase, CAZy, clan, superfamily

SUMMARY

Motivation: More than 1 % of genes in genomes code enzymes with glycosidase activities. On the basis of sequence similarity all known glycosidases have been classified into one hundred families. In many cases proteins of different families have a common evolution origin. It makes necessary to develop a hierarchical classification of glycosidase, which would reflect the degree of their relationship. The $(\beta/\alpha)_8$-barrel is the most common protein fold among glycosidases, that is why a classification of this group has the primary importance.

Results: Taken together, pairwise sequence comparison, analysis of the order of sequence appearance in the BLAST search results, and phylogenetic tree topology allow us to distinguish several subfamilies in a glycosidase family. Iterated BLAST screening and comparison of 3D structures allow to reveal relationships between members of different families. Based on our original results and published data of other authors, we have developed a hierarchical classification of the $(\beta/\alpha)_8$-barrel glycoside hydrolase catalytic domains.

INTRODUCTION

Glycoside hydrolases or glycosidases [EC 3.2.1.-] are a widespread group of enzymes of significant biochemical, medical, and industrial importance that hydrolyze the glycosidic bonds between two carbohydrates or between a carbohydrate and an aglycone moiety. A large multiplicity of these enzymes is a consequence of the extensive variety of their natural substrates: di-, oligo-, and polysaccharides. The traditional nomenclature of glycosidases (IUBMB) is based on their substrate specificity and occasionally on the molecular mechanism of their action; such a classification, however, does not reflect the structural features and evolutionary relationships of these enzymes, and it is not appropriate for enzymes that act on several substrates.

Accumulation of glycosidase sequences in databases has allowed to propose a sequence-based classification of glycoside hydrolases (Henrissat, 1991). This classification is regularly updated and now it is available at the CAZy site (http://www.cazy.org/CAZY/fam/acc_GH.html). It covers about 30,000 sequences of glycosidases and their homologues, which are grouped into more than 100 families. The molecular mechanism of their action (with inverting or retaining of the anomeric configuration) is conserved among members of a certain family. A careful examination of amino acid sequences and the tertiary protein structures allows to show the evolutionary

relationships among many glycosidase families. The relative families, having the same catalytic mechanism, were proposed to be grouped into clans (Henrissat, Bairoch, 1996). Currently, 14 clans (GH-A – GH-N) are described, and in total they contain 46 families. The largest of them (clan GH-A) includes 17 families; the others consist of 2 or 3 families each.

However, two different clans have never been merged in the CAZy classification, even when their significant similarity has been established. Moreover, relationships between glycosidases appear to be more complex: some glycosidases having different mechanisms of hydrolyzing reaction have been shown to be homologous. For example, we have described the furanosidase (or β-fructosidase) superfamily that includes clans GH-F (inverting glycosidases) and GH-J (retaining glycosidases), as well as the GHLP (COG2152 or DUF377) family of enzymatically-uncharacterized proteins (Naumoff, 2001).

Several CAZy families appear to present two drawbacks: (i) they are very large (up to 4,000 members in the case of GH13 family) and (ii) they contain enzymes that are poly-specific. Thus, it could be expedient to divide these large families into smaller clusters (= subfamilies) in order to better qualify the relationship between homologous glycosidases and better predict their substrate specificity. We have distinguished several subfamilies in the furanosidase superfamily (Naumoff, 2001).

The TIM-barrel, which has eight β/α motifs folded into a barrel structure, is one of the most usual protein folds (Nagano *et al*., 2001). A common origin of all $(\beta/\alpha)_8$-barrel domains has been proposed. According to this hypothesis, all of them have evolved from an ancestral $(\beta/\alpha)_4$-half-barrel by a tandem gene duplication, followed by a fusion and diversification (Lang *et al*., 2000). About 50 % of known glycosidases have $(\beta/\alpha)_8$-barrel fold of their catalytic domains (Rigden *et al*., 2003). They belong to four clans (GH-A, GH-D, GH-H, and GH-K), as well as to several other families, those have not been assigned to any clan.

In the present communication, we summarize data on the relationship of different $(\beta/\alpha)_8$-barrel glycosidases and propose for them a hierarchical classification.

METHODS

Protein sequences were retrieved from the NCBI database. Protein family analysis was performed using standard methods (Naumoff, 2006). Particularly, the phylogenetic trees were built using the Neighbor-Joining and Maximum Parsimony algorithms (PHYLIP package). Interfamily relationships were established by PSI-BLAST searches, using several most divergent representatives from each analyzed family. The number of iterations needed to reach a family member using the selected representative of a given family was considered as a degree of sequence similarity for two corresponding families. For the more details of interfamily comparison methods check our recent paper (Naumoff, 2005).

RESULTS AND DISCUSSION

Melibiases or α-galactosidases [E.C. 3.2.1.22] are glycosidases that cleave, with overall retention of the anomeric configuration, the terminal non-reducing α-D-galactose residues in α-D-galactosides, including galactose oligosaccharides, galactomannans, and galactolipids. The majority of known α-galactosidases have $(\beta/\alpha)_8$-barrel fold of the catalytic domain and belong to GH27 and GH36 families, which form clan GH-D. We have performed sequence analysis of all proteins from this clan. Subfamily structure was determined using pairwise sequence comparison (> 30 % of identity), analysis of the order of sequence appearance in the BLAST search results, and phylogenetic tree topology (monophyletic status). We distinguished three and four main subfamilies in GH27 and GH36 family, respectively.

Analysis of the PSI-BLAST search results suggests approximately the same evolutionary distance between GH27 and GH36 family proteins, and between them and members of GH31 family. This allows us to group these three families into the α-galactosidase superfamily. The phylogenetic analysis shows that GH27 and GH31 families are most probably monophyletic groups, but GH36 family has a polyphyletic origin. We propose to consider four subfamilies of GH36 family as four different families of glycosidases (GH36A-GH36D) within the α-galactosidase superfamily.

Iterated screening of the database by PSI-BLAST revealed distant sequence relationship of the α-galactosidase superfamily proteins with members of several other families, including GH13, GH97, and COG1649. GH13 is the largest family of glycoside hydrolases that contains enzymes of almost 30 different specificities comprising hydrolases, transferases, and isomerases. GH13 and two closely related families, GH70 and GH77, compose GH-H clan (or the α-glucosidase superfamily).

COG1649 (or DUF187) is a family of enzymatically-uncharacterized proteins. We distinguished four main subfamilies in this family. Iterated sequence analysis allowed us to reveal relationship of its members with proteins belonging to families GH13, GH20 (GH-K clan), GH31, and GH36. See the accompanying paper of A.Y. Kuznetsova and D.G. Naumoff for the more details of COG1649 analysis.

GH97 is a recently established glycoside hydrolase family that includes only two proteins with known enzymatic activity. We distinguished five main subfamilies in this family. Iterated BLAST searches showed the highest similarity of GH97 family proteins to members of the α-galactosidase superfamily. More distant relationship was found with GH20 and GH13 family glycosidases, as well as with some members of COG0535. COG0535 is a group of enzymatically-uncharacterized proteins. It was annotated as a family of predicted Fe-S oxidoreductases, like the closest COG0641 (http://www.ncbi.nlm.nih.gov/COG/). Our BLAST searches showed, that both COG families are related to the radical SAM superfamily of Fe-S enzymes, having (β/α)8-barrel fold.

3D-PSSM searches of the PDB database with several GH97 family proteins used as a query sequence yielded the highest level of 3D similarity with members of the α-galactosidase superfamily. Among other best hits we found representatives of several other glycoside hydrolase families: GH2 (GH-A clan), GH5 (GH-A), GH13 (GH-H), GH17 (GH-A), GH18 (GH-K), and GH20 (GH-K), as well as some other enzymes with $(\beta/\alpha)_8$-barrel fold.

Taken together, these data suggest a common evolutionary origin of glycosidase catalytic domains belonging to clans GH-A, GH-D, GH-H, and GH-K. Complementary results have been reported by several authors. Henrissat (1998) and Janeček (1998) found a distant sequence similarity of GH13 with GH31 and GH57 families, respectively. Imamura *et al*. (2001) proposed a common ancestor for family GH38 and GH57 glycoside hydrolases. Nagano *et al*. (2001) suggested an evolutionary relationship of GH-A clan with GH13 and GH14 families. A common origin of GH13, GH27, GH31, GH36, and GH66 families was proposed by Rigden (2002). Fold reconstruction methods allowed to predict a TIM-barrel type catalytic domain in GH29, GH44, GH50, GH71, GH84, GH85, and GH89 glycosidase families (Rigden *et al*., 2003). Moreover, GH50 family was suggested to be a common evolutionary ancestor of GH14 and GH42 families.

Based on the CAZy two-level classification and taking into consideration the data mentioned above, as well as protein classifications at SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) and Pfam (http://www.sanger.ac.uk/Software/Pfam/) sites, we propose a hierarchical classification of the TIM-barrel type glycoside hydrolases (Fig. 1).

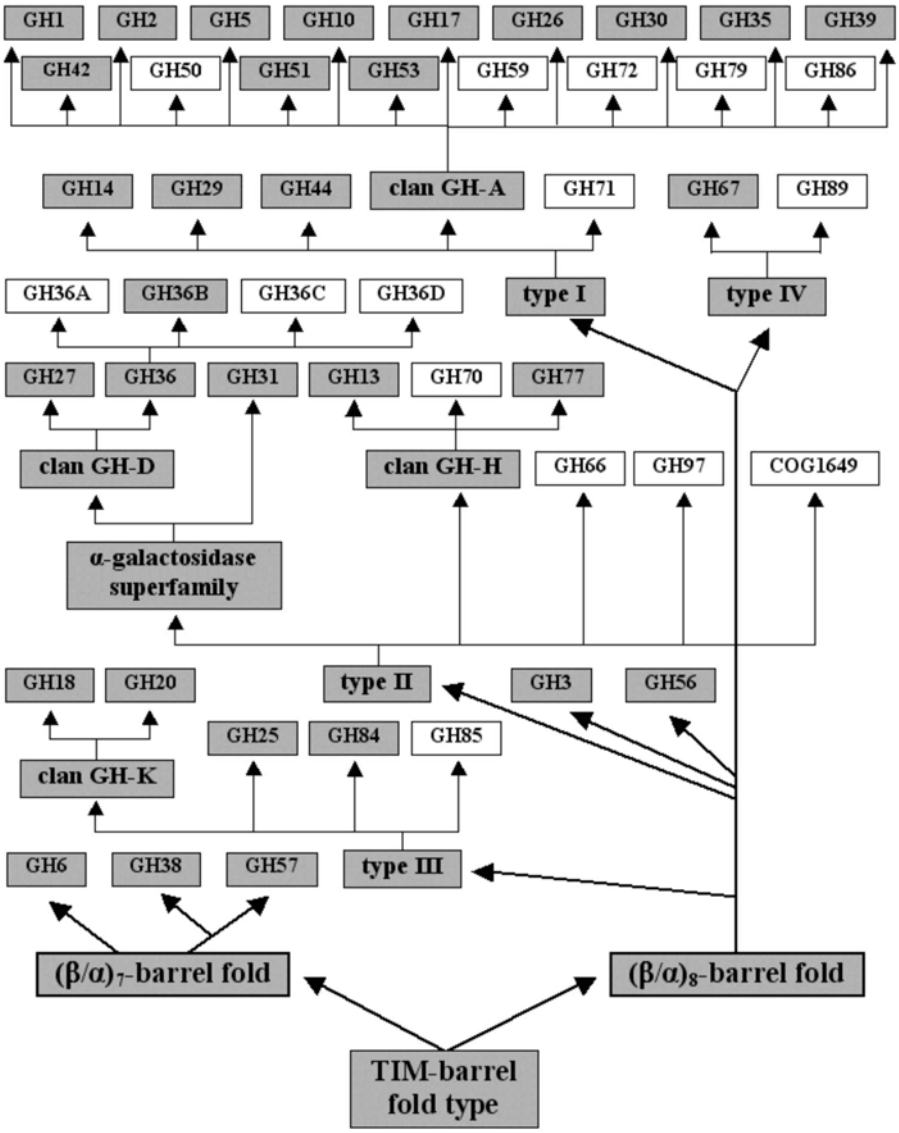


Figure 1. Hierarchical classification of the TIM-barrel type glycoside hydrolases proposed in this work. Subfamily structure is not shown. Shaded rectangles correspond to hierarchical groups that include at least one representative with known tertiary structure.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grant 06-04-49079-a) and by grants of the Russian President for young scientists (MK-118.2003.04 and MK-1461.2005.4).

REFERENCES

Henrissat B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.*, **280**, 309–316.

Henrissat B. (1998) Glycosidase families. *Biochem. Soc. Trans.*, **26**, 153–156.

Henrissat B., Bairoch A. (1996) Updating the sequence-based classification of glycosyl hydrolases. *Biochem. J.*, **316**, 695–696.

Imamura H., Fushinobu S., Jeon B.-S., Wakagi T., Matsuzawa H. (2001) Identification of the catalytic residue of *Thermococcus litoralis* 4-α-glucanotransferase through mechanism-based labeling. *Biochem.*, **40**, 12400–12406.

Janeček Š. (1998) Sequence of archaeal *Methanococcus jannaschii* α-amylase contains features of families 13 and 57 of glycosyl hydrolases: a trace of their common ancestor? *Folia Microbiol.*, **43**, 123–128.

Lang D., Thoma R., Henn-Sax M., Sterner R., Wilmanns M. (2000) Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.

Nagano N., Porter C.T., Thornton J.M. (2001) The $(\beta/\alpha)_8$ glycosidases: sequence and structure analyses suggest distant evolutionary relationships. *Protein Eng.*, **14**, 845–855.

Naumoff D.G. (2001) β-Fructosidase superfamily: homology with some α-L-arabinases and β-D-xylosidases. *Proteins*, **42**, 66–76.

Naumoff D.G. (2005) GH97 is a new family of glycoside hydrolases, which is related to the α-galactosidase superfamily. *BMC Genomics*, **6**, Art. 112.

Naumoff D.G. (2006) Phylogenetic analysis of a protein family. *Zbio*, **1**, Art. 3 (http://zbio.net/bio/001/003.html).

Rigden D.J. (2002) Iterative database searches demonstrate that glycoside hydrolase families 27, 31, 36 and 66 share a common evolutionary origin with family 13. *FEBS Lett*, **523**, 17–22.

Rigden D.J., Jedrzejas M.J., de Mello L.V. (2003) Identification and analysis of catalytic TIM barrel domains in seven further glycoside hydrolase families. *FEBS Lett.*, **544**, 103–111.

PROF_PAT: THE UPDATED DATABASE OF PROTEIN FAMILY PATTERNS. CURRENT STATUS

***Nizolenko L.Ph.*[*1], *Bachinsky A.G.*[1], *Yarygin A.A.*[1], *Naumochkin A.N.*[1], *Grigorovich D.A.*[2]**

[1] SRC VB "Vector", Koltsovo, Novosibirsk, 633159, Russia; [2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia, e-mail: odip@bionet.nsc.ru

[*] Corresponding author: e-mail: nizolenko@vector.nsc.ru

Key words: protein families, patterns, data banks, amino acid sequences, protein comparison

SUMMARY

Motivation: **Pro**tein **f**amily **pat**terns bank Prof_Pat is a collection of the patterns of groups of related proteins, characterizing the position intervals conservative in aligned proteins, and flexible fast search program.

Thus it is one of the numerous "secondary" banks, in which the information on the whole groups (families) of the related proteins, most typical and frequently unique features of this group is concentrated.

When "secondary" bank is constructed, the following properties are very important:

- the completeness of representation of prototype banks proteins,
- the sensitivity and specificity of the analysis,
- the actuality of the presented information.

These features do convenient application of a database for the protein analysis that in turn, is a necessary condition of database existence.

Results: We represent Prof_Pat as an updated, developing and improved tool for a prediction of function and distant similarity of proteins, satisfying all the listed conditions. Besides we offer variants of use of this database for the analysis of large groups of amino acid sequences.

Availability: http://wwwmgs2.bionet.nsc.ru/mgs/programs/prof_pat/ Prof_Pat local version is available via ftp: ftp://ftp.ebi.ac.uk/pub/databases/prof_pat/ and ftp://ftp.bionet.nsc.ru/pub/biology/vector/prof_pat/.

INTRODUCTION

Now alongside with amino acid sequences databases, the general recognition and a wide circulation have received so-called "secondary" databases. These bases are used for the analysis of amino acid sequences with the purpose of a prediction of functions and related communications of coded proteins. Bank Prof_Pat created and supported in the SRC VB "Vector" is a "secondary" database too.

METHODS AND ALGORITHMS

Protein family patterns, the bank of these patterns Prof_Pat and flexible fast search program were created using original technology (Bachinsky *et al*., 2000). The version of

Prof_Pat 1.18, constructed on the basis of the UniProt 6th release, including the 48th release of the Swiss-Prot bank and 31th release of TrEMBL (Wu *et al*., 2006), contains patterns of 138788 groups of related proteins including more than 1000000 amino acid sequences.

IMPLEMENTATION AND RESULTS

Prof_Pat was earlier showed to have as good completeness and variety of included proteins as the best world-known "secondary" banks. At the same time, its specificity and sensitivity is higher than those of other banks, and its search speed was 3–10 times higher. (Nizolenko *et al*., 2003).

For example, from the 920.402 Swiss-Prot (rel. 42) and TrEMBL (rel. 25) sequences, that have the reference of Interpro, one of the largest and modern banks (Mulder *et al*., 2003), in their description, only 4 ones were no recognised by Prof_Pat (rel.1.14) patterns. At the same time, 14185 sequences, that have no Interpro reference as well as any detailed description of a putative function for the protein, show very good similarity with well-described Prof_Pat patterns. (Nizolenko *et al*., 2005b).

As databases of primary amino acid sequences are continuously replenished, updating of "secondary" banks is very important.

Prof_Pat is constantly updated database. The information on growth of this secondary base in parallel to growth of Swiss_Prot /TrEMBL volumes is submitted in Table 1.

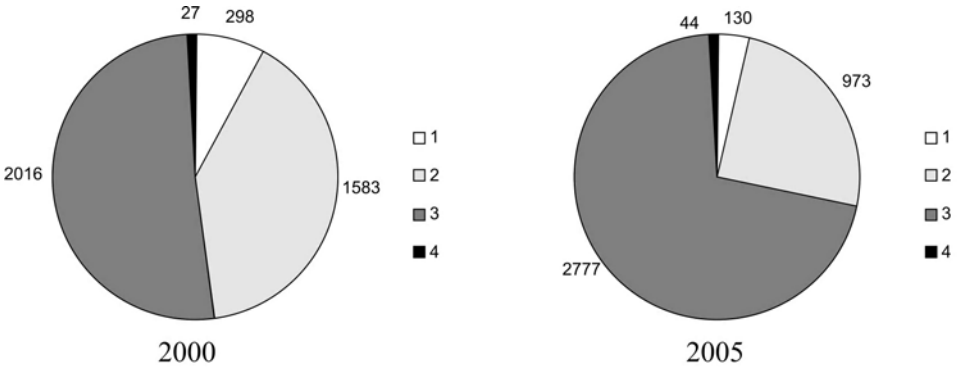
Table 1. The growth and development of Prof_Pat bank in parallel to growth of Swiss-Prot and TrEMBL volumes

Prof_Pat release/data	Swiss-Prot /TrEMBL release	Full-length sequences in Swiss-Prot/ TrEMBL	Sequences in Prof_Pat	Patterns in Prof_Pat	Proportion of the in Swiss-Prot/ TrEMBL full-length sequences having hits to Prof_Pat
1.1 1998-99	29/1	~98000	52122	7083	0.57
1.6 Oct 2000	39/15	295932	166667	24692	0.59
1.7 Apr 2001	39/16	320511	181644	27187	0.65
1.8 Nov 2001	40/17,18	385437	217360	31613	0.66
1.10 May 2002	40/20	475343	283765	41076	0.69
1.11 Jan 2003	40/21,22	556538	344429	50149	0.70
1.14 Jan 2004	42/25	784262	509506	71619	0.74
1.16 Dec 2004	44/27	1010596	676485	90506	0.76
1.17 Jul 2005	46/29	1219335	822781	106725	0.77
1.18 Apr 2006	48/31	1634672	1084331	138788	0.79

The completeness of representation of the prototype bank proteins and ability to distinguish sequences which have not entered into Pof_Pat are growing too. More than 79 % of the full-length sequences in Swiss-Prot 48 + TrEMBL 31 have at least one hit to Prof_Pat patterns, whereas in 1998 - less then 60 %.

Growth and updating of the database is accompanied by improvement of quality of predictions. The investigation of the amino acid sequences of open reading frames of the complete genome of *Mycobacterium tuberculosis* using the protein family pattern bank Prof_Pat carried in 2005 in comparison with 2000 is presented in Fig. 1.

Salmonella typhi strain CT18 can be another example of successful Prof_Pat investigation of amino acid sequences, translated from complete genome. Of 4395 open reading frames of this microorganism, possible function or similarity with hypothetical proteins family were predicted by Prof_Pat 1.16 for 4246 (more then 96 %) with high significance level.



1. No detected similarity with any Prof_Pat proteins group.
2. Sequences recognised by Prof_Pat with low similarity level.
3. The similarity to proteins, described by other investigators (Cole S.T. et al. 1998) and confirmed by Prof_Pat.
4. New similarity. This protein's functions were not predicted up to now. For 35 of them Prof_Pat 2005 results are confirmed by investigation with other database such as Pfam, Interpro (Nizolenko et al.2005a).

Figure 1. Investigation of *Mycobacterium tuberculosis* amino acid sequences with Protein Family Patterns Bank Prof_Pat carried in 2005 in comparison with 2000.

Not only the quantity of distinguished sequences was increased. The results of predictions became more authentic, they are confirmed by investigation with other database such as Pfam and Interpro (Nizolenko *et al*., 2005a).

CONCLUSION

Expediency of use of bank Prof_Pat as an updated, developing and improved tool for a prediction of function and relationships of proteins was demonstrated again. It allows to get new information for assumption of structural and functional similarity for distinct proteins as well as for large groups of amino acid sequences.

REFERENCES

Bachinsky A.G. *et al*. (2000) PROF_PAT 1.3: updated database of patterns used to detect local similarities. *Bioinformatics*, **16**, 358–366.

Cole S.T. *et al*. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.

Mulder N.J. *et al*. (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucl. Acids Res*. **31**, 315–318.

Nizolenko L.Ph. *et al*. (2003) Database of patterns PROF_PAT for detecting local similarities. *In Silico Biology*, **3**, 205–213.

Nizolenko L.Ph. *et al*. (2005a) Study of the amino acid sequences of open reading frames of the complete genome of Mycobacterium tuberculosis using the protein family pattern bank Prof_Pat. *Biofizika*, **50,** 986–992. (In Russ.).

Nizolenko L.Ph. *et al*. (2005b) Database of patterns for detecting local similarities PROF_PAT in 2005. *Proceeding of the International Moscow Conference on Computational Molecular Biology* (MCCMB), 18-21 July. 2005, 253–255.

Wu C.H. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucl. Acids Res.*, **34**, Database issue D187–D191.

ACTION OF MEMBRANE-ACTIVE PEPTIDES ON EXPLICIT LIPID BILAYERS. ROLE OF SPECIFIC PEPTIDE-LIPID INTERACTIONS IN MEMBRANE DESTABILIZATION

***Polyansky A.A.*[*1, 2], *Aliper E.T.*[1, 2], *Volynsky P.E.*[1], *Efremov R.G.*[1]**
[1] M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, 117997, Russia; [2] Biological Department, M.V. Lomonosov Moscow State University, Moscow, 119992, Russia
[*] Corresponding author: e-mail: newant@nmr.ru

Key words: molecular dynamics, explicit membrane model, peptide-membrane interactions

SUMMARY

Motivation: Membrane-active peptides play a crucial role in numerous cell processes, such as fusion, transport of therapeutic compounds, disturbance of integrity of membranes, and others. Many of them act as highly specific and efficient drugs and, therefore, attract growing interest for biomedical applications. Because of experimental difficulties with characterization of their spatial structure and mode of membrane binding, essential attention is given now to molecular modeling techniques.

*Results:*The present work sums up our recent results on molecular dynamics (MD) simulations of binding of several membrane active (fusogenic, antimicrobial and cell penetrating) peptides (MAP) to hydrated lipid bilayers differing in length, saturation degree of acyl chains, chemical nature and/or charge of headgroups. The character of "membrane response" may significantly vary depending on the peptide type and it target membranes. However, a number of common phenomena in MAP-membrane interactions can be outlined. The mode of peptide binding to membrane correlates with a rate of membrane destabilization – deeper insertion of a peptide leads to the more prominent effect. The "membrane response" induced by a peptide may be caused by formation of specific peptide-lipid complexes. Relationships between character of interactions of peptides with membranes and the mechanism of their action are also discussed.

INTRODUCTION

Membrane-active peptides (MAP) represent a large class of compounds from either natural or synthetic source which possess a wide range of membrane activity often concerned with alteration of properties of the host membranes. It is significant that the functioning of such peripherally bound peptides is mediated by a number of factors: their conformation, the mode of membrane binding, pH, the phase condition and lipid composition of membrane.

Another interesting aspect of MAPs' action is specific reorganization of structural and dynamic properties of membrane ("membrane response") induced by peptide insertion. Therefore, the understanding of the structure-function relationship for MAP represents an intriguing challenge in the field of structural biology. Apart from fundamental importance (studies of general principles of protein insertion, folding and stabilization in bilayer), solving the problem is invaluable in the optimization of these molecules' behavior for

pharmaceutical and biotechnological applications, such as the development and targeted delivery of drugs through membranes, the design of MAP with prescribed properties, gene therapy, and disease control. Unfortunately, studies of biological membrane-protein systems are very difficult because of their complexity.

Possible solution of this problem is employment of different membrane mimic systems such as micelles of detergents, lipid vesicles and bilayers. However, experimental techniques often give only overall picture of behavior of a model peptide-membrane system, while in many cases molecular details of peptide-membrane interactions are missing. The development of molecular modeling approaches would be indispensable in avoiding these problems. Such methods have begun to be widely used in studies of protein-membrane interactions. Among them molecular dynamics (MD) of peptides and proteins in explicit membrane environments (lipid bilayers and micelles) is the most powerful one. This method permits investigation of the actions of membrane-active agents on the atomic level, including microscopic details of their interactions with lipids and detergents, changes in peptide structure during binding and specific "membrane response" induced by peptide insertion.

Here, we present comparative MD simulations studies of binding of several MAPs to hydrated lipid bilayers. Among them are fusogenic (FP), antimicrobial (AMP) and cell penetrating peptides (CPP). The emphasis is made on detailed characterization of peptide-lipid interactions and their role in membrane destabilization. To address this question MD simulations were performed in pure bilayers differing in length, saturation degree of acyl chains, chemical nature and/or charge of headgroups and lipid mixtures close in composition to real cell membranes (Table 1).

Table 1. Simulated MAPs

Peptide	Type	Sequence	Initial conformation	Bilayers
Latarcin 2a (Ltc2a)	AMP	GLFGKLIKKFGRKAISY AVKKARGKH	α-helical, NMR in SDS micelles (2G9P)	$POPE_{200}$:$POPG_{88}$, $POPC_{114}$:$POPE_{114}$: $CHOL_{60}$
E5	FP	GLFEAIAEFIEGGWEGL IEG	α-helical, NMR in DPC micelles	$DMPC_{128}$, $DPPC_{128}$
Penetratin (pAntp)	CPP	RQIKIWFQNRRMKWKK	α-helical, NMR in water/TFE mixture (1KZ0)	$DOPS_{128}$, $DOPC_{128}$

METHODS

Systems preparation. The model bilayers were constructed in such a way that their structural parameters (area per lipid molecule, bilayers thickness) corresponded to experimental data. The compositions of mixed bilayers were similar to those in cell membranes (human erythrocyte, *E. coli*). The systems were solvated by SPC water molecules and neutralized by adding the necessary counterions. All systems were subjected to energy relaxation via 5×10^4 steps of steepest descent minimization followed by heating from 5 K to the temperature of simulations during 50-ps MD run. Then the long-term collection MD runs were carried out. The final configurations of bilayers obtained in these calculations were further used in MD simulations of their complexes with peptides.

Simulation details. All simulations were performed using the GROMACS 3.2.1 (Lindahl *et al*., 2001) package and the GROMOS87 force field specially adopted for lipids (Van Gunsteren *et al*., 1987). Simulations were carried out with a time step of 2 fs, with imposed 3D periodic boundary conditions, in the NPT ensemble with isotropic pressure of 1 bar. Van der Waals interactions were truncated using the twin range 12/20 Å spherical cutoff function. Electrostatic interactions were treated in two different ways: by

use the same cutoff scheme and the PME algorithm (1.2 Å Fourier spacing). Previously it has been shown that the cutoff-based MD simulations of explicit bilayers may lead to serious computational artifacts only in a case of electrostatically heterogeneous systems like charged lipids and counterions (Polyansky *et al*., 2005). That is why to save a CPU-time, MD simulations of zwitterionic membrane systems were carried out with the cutoff functions for electrostatic interactions. All components of the systems were coupled separately using the Berendsen thermostat to a temperature bath with a coupling constant of 0.1 ps. The simulation temperatures were chosen to ensure the liquid crystalline phase of lipid bilayers.

Analysis of MD-trajectories. Analysis of MD trajectories was performed using original software developed by the authors and utilities supplied with the GROMACS package. The following parameters of the simulated systems were analyzed: bilayer structure (the area per lipid molecule (A_L), the bilayer thickness (D_{PP}), etc); peptide secondary structure; peptide-lipid interactions (energies of electrostatic and van der Waals interactions, long-living peptide-lipid contacts); 2D distribution of various structural and dynamic properties of bilayers. All these parameters were averaged over the equilibrium parts of corresponding MD trajectories.

RESULTS AND DISCUSSION

Long term simulations (~20–30 ns) of considered peptide-membrane systems permit delineation several similar effects of MAPs' interactions with model bilayers. First, all peptides undergo the reorganization of their initial secondary structures. Thus, unlike the aqueous solution, the water-membrane interface significantly promotes structuring of the anchored amphiphilic peptide. It is interesting, that this effect is more prominent in a case of strongly amphiphilic molecules, like Ltc2a and E5, than in a case of slightly amphiphilic pAntp. Second, analysis of the accumulated MD trajectories reveals no essential alterations in global structural properties of the bilayers (A_L, D_{pp}, etc) – the peptide-induced destabilization of the bilayer structure has a local character. Essential insight into the microscopic picture of the "membrane response" may be gained *via* detailed analysis of in-plane distributions of various bilayer characteristics (dynamical lipid fractions (see Fig. 1*b*), angles of lipid tails and heads to bilayer normal, surface hydrophobic/hydrophilic properties).

These results reveal that peptides under study possess a prominent effect of destabilization of model membranes. The character of "membrane response" may significantly vary depending on the peptide type and it target membranes. However, a number of common phenomena in MAP-membrane interactions can be outlined. The mode of peptide binding to membrane correlates with a rate of membrane destabilization – deeper insertion of a peptide leads to the more prominent effect. During MD simulations peptides form long-living complexes with lipid heads (i.e. several peptide residues are associated with lipid head). Structure of such complexes depends on the peptide's conformation, its amino acid composition and chemical nature of lipid heads (see Fig. 1*c*). Lipids associated with peptide significantly differ in structural and dynamic properties compared with the rest of the bilayer.

Thus, the "membrane response" induced by the peptide may be caused by formation of specific peptide-lipid complexes. These observations open up novel intriguing opportunities for *de novo* design of MAPs with prescribed character of membrane destabilizations and/or selectivity to different target membranes.

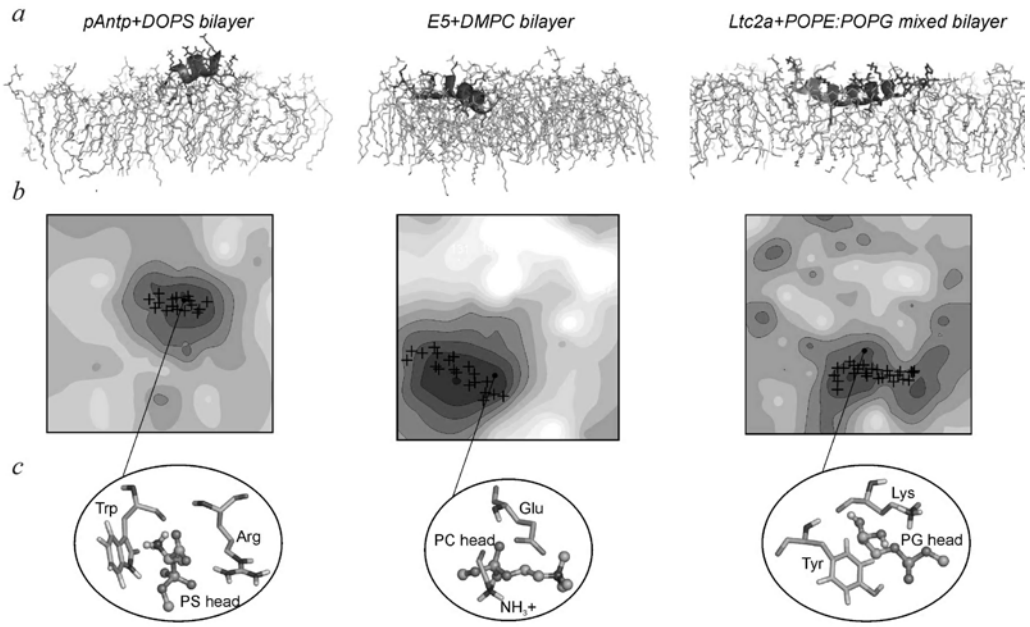


Figure 1. Interactions of peptides with model bilayers. *a*– binding mode of peptide. Lipids are shown with sticks, peptides are given in ribbon representation. *b* – 2D distribution of fluctuations (RMSF) of lipid heads. Dark regions correspond to low RMSF values. The peptide backbones are shown with black crosses. *c* – examples of long-living complexes of peptide residues with lipid heads. Peptide residues and lipid heads are given in "stick" and "ball and stick" representations, respectively.

ACKNOWLEDGEMENTS

This work was supported by the Programme RAS MCB, by the Russian Foundation for Basic Research (grant No. 04-04-48875-a), by the Russian Federation Federal Agency for Science and Innovations (The State contract 02.467.11.3003 of 20.04.2005, grant SS-4728.2006.4). We thank Dr. Dubovskii P.V. for the provided NMR-structures of Ltc2a and E5.

REFERENCES

Lindahl E., Hess B., Spoel D. van der. (2001) *J. Mol. Mod.,* **7**, 306.

Polyansky A.A., Volynsky P.E., Nolde D.E., Arseniev A.S., Efremov R.G. (2005) *J. Phys. Chem. B*, **109**(31), 15052.

Van Gunsteren W. F., Berendsen H.J.C. (1987) *Gromos-87 manual*. Biomos B.V. Nijenborgh 4, 9747 AG Groningen, the Netherlands.

COMBINING MOLECULAR DOCKING WITH RECEPTOR DOMAIN MOTIONS: SIMULATIONS OF BINDING OF ATP TO CA-ATPase

***Pyrkov T.V.*[*1, 2], *Kosinsky Yu.A.*[1], *Arseniev A.S.*[1], *Priestle J.P.*[3], *Jacoby E.*[3], *Efremov R.G.*[1]**

[1] M.M. Shemyakin and Yu. A. Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russia; [2] Moscow Institute of Physics and Technology, Moscow, Russia; [3] Novartis Institutes for Biomedical Research, Basel, Swizerland

[*] Corresponding author: e-mail: pyrkov@nmr.ru

Key words: receptor flexibility, molecular dynamics, ATP-binding

SUMMARY

Motivation: Most molecular docking algorithms consider only flexibility of ligand molecules while receptor is held rigid. At the same time it becomes evident that receptor's flexibility is indispensable for obtaining correct structures of protein-ligand complexes via docking simulations.

Results: To study the effect of receptor flexibility we used the ATP – Ca-ATPase complex. In experimental structures of the complex ATP-analogues simultaneously interact with two active sites which are situated in different protein domains. In the apo-form the distance between these sites is too large and standard docking protocols fail to reproduce the experimental structure. We used molecular dynamics simulations of nucleotide-binding part of Ca-ATPase to study its domain motions and generate a representative ensemble of the receptor conformations. This approach allowed correct prediction of the structure of the complex based on the apo-structure of the Ca-ATPase.

Availability: http://model.nmr.ru.

INTRODUCTION

Most docking algorithms incorporate ligand flexibility while keeping the protein rigid. But structural studies of ligand-protein associations indicate that these processes are often accompanied by conformational changes in protein structure. In many cases these changes are related to the protein side-chain conformations although main-chain motions of flexible loops are also observed. Moreover, for some proteins (e.g. ATP – Ca-ATPase complex (Toyoshima *et al*., 2004)) changes in receptor structure involve even relative domain motions. Such effects indicate the necessity of modeling the receptor's flexibility in the docking procedure.

Nowadays a lot of various approaches are applied to incorporate the flexibility of receptor into the docking algorithm. As a rule, all of them are aimed at collecting a representative ensemble of protein conformations. After this is done the standard docking algorithm of a flexible ligand to rigid receptor is performed for each of them. The ways of generating such ensemble may vary – from collecting different X-ray structures of the target protein and stochastic exploring its conformational space (sampling method based on bonds network, "rotamer libraries") to taking snapshots of molecular dynamics (MD) simulations. However, the latter approach yet has very few examples (Cavasotto *et al*.,

2005) of applications to investigate the influence of global domain motions on docking simulations. Our work is dedicated to implementation of MD simulations to sample global domain rearrangements and local side-chain motions of ATP-binding part of Ca-ATPase as a target for ATP docking.

METHODS

MD of ATP-binding domain of Ca-ATPase was performed for 3 experimental models: 1EUL ("open" apo-form), 1IWO (another "open" apo-form), and 1T5S ("closed" holo-form). The GROMACS program and the GROMOS96 force field (Berendsen et al., 1995) were used. Molecules with uncharged N- and C-termini were placed in rectangular boxes of a simple point charge (SPC) model of water molecules with edges of 10 Å. The three-dimensional periodic boundary conditions were imposed. To relax the system we performed 300 steepest descent iterations with fixed protein atoms followed by 300 conjugate-gradients steps with fixed backbone, and finally 300 conjugate gradients steps without constraints. Then the system was subjected to a 10 ps MD run in an NPT (constant pressure and temperature) ensemble with fixed protein atoms. Finally, it was heated from 5 to 300 K for 60 ps in an NVT (constant volume and temperature) ensemble. After that the production run was performed in an NPT ensemble at 300 K. Non-bonded interactions were truncated with the twin-range cutoff of 10/18 Å. The length of each MD trajectory exceeded 2 ns so as to allow detection of domain rearrangements. To characterize the relative domain orientation we introduced the interdomain angle Θ, determined as the angle between two vectors originating from the centre of the hinge region and pointing to the centres of the N- and P-domains. For the ATP docking experiments the snapshots of protein structure from the first 2 ns of each MD run were taken at 8 ps interval.

To perform docking of ATP we used 8 high-resolution experimental structures of Ca-ATPase available from the Protein Data Bank and a set of conformations from MD trajectories. To dock ATP into the ATP-binding domain of Ca-ATPase we used the GOLD 2.0 (Jones *et al.*, 1997) software with the scoring function "goldscore". The docking sphere radius was 21 Å, the origin of this sphere was atom CD1 of Phe482. 60 runs of the docking procedure were performed for each receptor conformation. All ligands and water molecules were removed from the receptor structure prior to docking. To retrieve a correct conformation of ATP from a diverse set produced by docking we used our special ATP-oriented ranking criterion. It is based on the analysis of ATP-protein interactions in a set of experimental structures of ATP-protein complexes. This criterion is based mainly on hydrophobic interactions of the adenine moiety of ATP with its protein environment.

RESULTS AND DISCUSSION

The conformation of the ligand in the N-domain was reproduced in all receptor structures crystallized with ATP-analogs (Table 1). Also, ATP conformation close to the experimental one was found for the apo-form 1IWO. However, docking failed to predict ATP poses in the N-domain for other apo-forms of the receptor. This indicates that docking may yield unsatisfactory results if the site geometry is not optimized for a particular ligand. Moreover, in the experimental models bound ATP-analogues interact simultaneously with active sites in N and P-domains. Mutual disposition of the latter ones corresponds to the interdomain angle Θ ~111°. This was reproduced by docking for the receptor structures taken from complexes with ATP-analogues. But in the "open" forms of Ca-ATPase (corresponding to the values of Θ ~125°-165°) the distance between these sites is approximately twice as large as the size of ATP molecule. That means that in such cases standard docking algorithms which consider receptor rigid do not allow successful prediction of ligand-receptor complex structure.

Table 1. Results ATP docking into experimental models of Ca-ATPase

PDB code	Ligand	Angle Θ	Rank of correct docking solution	
			"Goldscore"	ATP-criterion
1VFP	ACP	112°	1	1
1T5T	ADP	111°	1	1
1T5S*	ACP	111°	7	1
1WPE	ADP	111°	1	1
1WPG	ADP	157°	20	1
1IWO*	–	125°	33	11
1XP5	–	155°	–	–
1EUL*	–	165°	–	–

* For these models MD simulations were performed.

To take into account flexibility of the receptor we used MD simulations to generate an ensemble of its conformational states starting from 3 different experimental models. In the calculated MD trajectories high-amplitude relative domain motions of type "closure" were detected which corresponds to available experimental structures of Ca-ATPase.

For each trajectory it was shown that it is possible to obtain correct docking solutions for the N-domain site for nearly a half of the receptor conformations. Also it is possible to dock ATP to both active sites when the relative domain orientation is appropriate, namely it was found that such docking solutions can be obtained for MD-conformations, with the angle $\Theta < 120°$. Such "closed" form was observed in each trajectory independently of the starting conformation.

Proper scoring of the docking solutions needs a special consideration. Thus, docking yielded correct solutions for 6 experimental Ca-ATPase structures, but ranked them top only in 3 cases. To improve ranking we applied our ATP-criterion which ranked correct solution top in 5 cases and improved the rank of such solution in the remaining case. The results over all ATP poses for each MD trajectory were similar. Comparison with the experimental structure of the ATP – Ca-ATPase complex showed that among ~14000 solutions only ~450 were correct predictions. The distribution of these correct solutions when ranked by ATP-criterion is considerably better than that ranked by the "goldscore" function, which is close to a random distribution (Fig. 1).

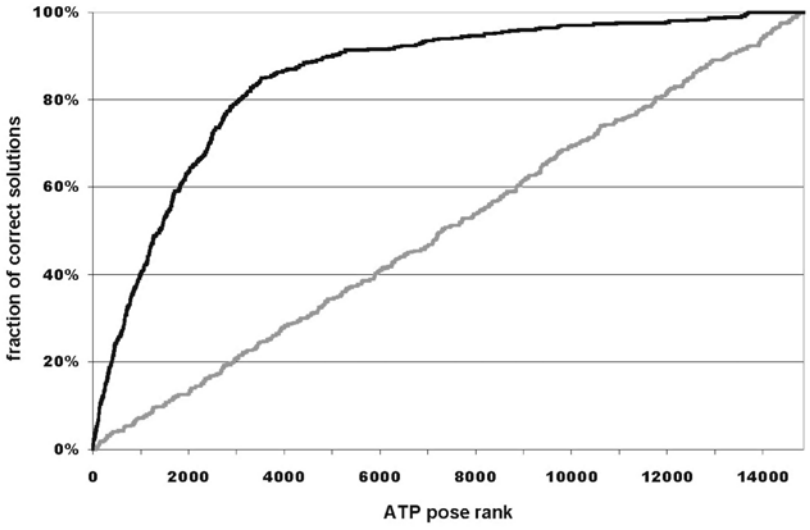


Figure 1. Ranking of docking solutions by the "goldscore" function (gray) and by ATP-criterion (black) over all MD-conformations of the receptor (here are shown the results for 1EUL). The fraction of correct docking solutions (~ 450 in totals) is shown.

ACKNOWLEDGEMENTS

This work was supported by the Russian Foundation for Basic Research (grants Nos 04-04-48875-a, 05-04-49283-a, 06-04-49194-a) and by the Russian Federation Federal Agency for Science and Innovations (The State contract 02.467.11.3003 of 20.04.2005, grant SS-4728.2006.4).

REFERENCES

Berendsen H.J.C., van der Spoel D., Drunen R. (1995) GROMACS. *Comp. Phys. Comm.*, **91**, 43–56.

Cavasotto C.N., Orry A.J.W., Abagyan R.A. (2005) The challenge of considering receptor flexibility in ligand docking and virtual screening. *Curr. Comput. Aided Drug Des.*, **1**, 423–440.

Jones G., Willet P., Glen R.C., Leach A.R., Taylor R. (1997) Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, **267**, 727–748.

Toyoshima C., Nomura H., Tsuda T. (2004) Lumenal gating mechanism revealed in calcium pump crystal structures with phosphate analogues. *Nature*, **432**, 361–368.

HOW ARE CHARGED RESIDUES DISTRIBUTED AMONG FUNCTIONALLY DISTINCT STRUCTURAL DOMAINS OF AMINOACYL-tRNA SYNTHETASES?

***Safro M.*[*1], *Tworowski D.*,[1] *Feldman A.*[2]**
[1] Department of Structural Biology, Weizmann Institute of Science, 76100, Rehovot, Israel;
[2] EML Research gGmbH, Schloss-Wolfsbrunnenweg 31c, 69118 Heidelberg, Germany
[*] Corresponding author: e-mail: mark.safro@weizmann.a.il

Key words: aminoacyl-tRNA synthetases, domain organization, primary structure, electrostatic potential, Poisson-Boltzmann theory

SUMMARY

Motivation: Unlike many other families of enzymes, which catalyze the same overall reaction, aminoacyl-tRNA synthetases (aaRSs) are extremely heterogeneous in terms of primary sequence and subunit organization. For the most part aaRSs are negatively charged at physiological conditions, as are tRNA substrates. What are the driving forces that ensure an attraction between like-charged macromolecules? As may be inferred from multiple sequence alignments (MSA), concentration of the invariant charged residues in structural domains doesn't correlate with contribution of the domains to formation of the electrostatic field at long distances.

Results: In aaRSs family the subset of evolutionary non-conserved charged residues generates long-range electrostatic potential (EP) similar to the native one. We evaluate contribution of individual structural domains to the EP generated by native (NS), conservative (CS) and non-conservative subsets (NCS) of charged residues. For monomeric IleRS and heterodimeric PheRS we further analyzed the interplay between the domain functionality and their role in the field formation at long distances.

INTRODUCTION

AaRSs are of primary importance in the transformation of the genetic information from mRNA into polypeptide chain covalently attaching appropriate amino acids to the corresponding nucleic acid adaptor molecules – tRNA via the two-step aminoacylation reaction. The attachment of the correct amino acid to a tRNA is the crucial step determining the accuracy of protein biosynthesis. AaRSs exist as monomers, α_2-dimers or tetramers of α_4 and $(\alpha\beta)_2$ types. Previously (Tworowski, Safro, 2003; Tworowski *et al.*, 2005), we evaluated the contribution of electrostatic interactions to formation of aaRS-tRNA encounter complexes. It has been shown that 3D-isopotential surfaces (IPS) generated by monomeric, dimeric and heterotetrameric synthetases at 0.01kT/e contour level reveal the presence of large positive patches ("blue spaces"), one for each tRNA substrate molecule. It is apparent that this characteristic landscape of aaRS's electrostatic potential is triggered by specific distribution of the charged residues along sequences and, thus in space.

A challenging problem is to identify the charge distribution along the polypeptide chain that substantially affects the topology of aaRS's ES field. Based on MSA, we subdivided the entire pool of aaRS's charged residues into three subsets: native (NS),

conservative (CS) and non-conservative (NCS). It is of interest that the total charge of CS is close to zero, whereas that of NCS is similar to the aaRS's net charge. According to Smoluchowski's theory of bimolecular association, the capture distance for two macromolecules is a sum of reactants hydrodynamic radii (Berg, von Hippel, 1985). Thus, each aaRS can be conceived as a reactive sphere (RES) built around the geometric center of enzyme with radius equal to the capture distance (R_{RES}). Our calculations made apparent the resemblance of shape and topography of positive patches at ±0.01 kT/e formed by NS and NCS, and essential differences in landscapes generated by the CSs.

METHODS

Multiple sequence alignment (MSA). ClustalW program (www.ebi.ac.uk/clustalw) with BLOSUM62 matrix and gap penalty of –12 was used. The sequences of different bacterial organisms were included in the MSA for each tested aaRS system. The sequences were derived from the Swiss-Prot Database (www.ebi.ac.uk/swissprot) and the Protein Data Bank (PDB) (www.rcsb.org/pdb). "Conservative subset" (*CS*) of charged residues includes strictly conserved residues as well as all meaningful substitutions (Asp ↔ Glu or Arg ↔ Lys) identified by MSA; "non-conservative subset" (*NCS*) consists of non-conserved charged residues.

Statistical analysis of charge distribution among aaRSs' domains. Domains' "charging" (D_i^{chrg}) and fractions of conserved (F_i^{CS}) and non-conserved (F_i^{NCS}) charged residues associated with itch domain are presented by:

$$D_i^{chrg} = N_i^{chrg} / N_{chrg} \quad (1), \qquad F_i^{CS} = N_i^{CS} / N_{CS} \quad (2), \qquad F_i^{NCS} = N_i^{NCS} / N_{NCS}, \quad (3)$$

where N_i^{CS}, N_i^{NCS} and N_i^{chrg} are the numbers of conserved, non-conserved and all charged residues of the *i*tch domain, respectively; N_{CS}, N_{NCS} and N_{chrg} are the conserved, non-conserved and total number of charged residues in aaRS's polypeptide chain.

Calculation of electrostatic potentials. Electrostatic potentials were calculated at each grid point on the reactive encounter sphere (RES), built around the geometric center of molecule, by using Poisson-Boltzmann equation implemented in Delphi4.0 (Rocchia *et al*., 2004). The standard ionization states of the charged residues at physiological pH 7 were applied, i.e. neutral form for the His and ionized forms for Asp, Glu, Arg and Lys. To model charge distribution of native subsets, the formal charges were assigned to all charged residues of the protein. To evaluate the contribution of individual domain to the electrostatic potential the original PDB-files were modified by switching off the charged residues of the domain.

Similarity analysis for electrostatic potentials. Electrostatic potentials φ for *in silico* modified aaRSs were calculated at the same grid points on RES, as native ones. For each pair of potentials (φ_{mod} and φ_{native}) in the native "blue space" area (i.e. at the points $N_{\varphi(+)}^{native}$), the Hodgkin similarity index (SI) was calculated (Blomberg *et al*., 1999):

$$SI = \frac{2(\vec{\varphi}_{\text{mod}}, \vec{\varphi}_{native})}{(\vec{\varphi}_{\text{mod}}, \vec{\varphi}_{\text{mod}}) + (\vec{\varphi}_{native}, \vec{\varphi}_{native})} \quad (4)$$

$$(\vec{\varphi}_{\text{mod}}, \vec{\varphi}_{native}) = \sum_{x,y,z} \varphi_{\text{mod}}(x,y,z)\varphi_{native}(x,y,z) \quad (5)$$

Here $(\vec{\varphi}_{\text{mod}}, \vec{\varphi}_{native})$, $(\vec{\varphi}_{\text{mod}}, \vec{\varphi}_{\text{mod}})$ and $(\vec{\varphi}_{native}, \vec{\varphi}_{native})$ denote scalar product calculated for all points within the region of positive patch; *x*, *y*, *z* are Cartesian coordinates of the grid points on RES.

RESULTS AND DISCUSSION: A CASE STUDY

By way of illustration we selected two types of aaRSs with different subunit organization: monomeric IleRS [α; (PDB code 1qu2)], and hetero-tetrameric PheRS [$(\alpha\beta)_2$; PDB code 1eiy].

The Hodgkin index was used as a measure of similarity between the native electrostatic potential and those produced by different subsets of charged residues. SI falls in the range from 1 to -1. The SI close to 1 indicates to a high degree of similarity, whereas 0 and -1 correspond to fully uncorrelated and anti-correlated potentials, respectively. As it follows from our results SI reaches its peak ~1 when electrostatic potential on RES is generated by NCS. In contrast, the electrostatic potential generated by CSs uncorrelated (SI ~ 0) with those of native set and NCS subset.

Analysis of CS's residues, distributed among aaRS's domains, reveals that larger portion of conserved charged residues is concentrated in catalytic domains (Fig. 1*a*, *c*). The Rossmann fold of IleRS contains 64 % of CS's residues while the catalytic domain of PheRS 46 %. However there are structural domains such as N-term, Cp2, Zn-binding of IleRS and B4 domain of PheRS in which conserved charged residues were not found. This suggests dissimilar distribution of CS's residues among different aaRS's domains.

The intriguing result is that, regardless domain's charging, the switching-off charged residues from certain domains has no significant impact on the distribution of electrostatic potential on RES. It is of interest that contribution of some domains to the EP on RES remains small, even though the concentration of non-conserved charged residues (F_i^{NCS}) there is relatively high. Thus high degree of similarity to native EP is observed for IleRS, when contribution of non-conserved charged residues from the N-terminal, Rossmann-fold, Cp2, Helical, C-term junction or Zn-binding domains is alternately excluded. This is evidenced by proximity of SI values to 1 (Fig. 1*b*). In case of PheRS, when charged residues are switched off within the domains B1, B3-B8 or catalytic module, the resulting electrostatic potentials are remain unchanged in compare to the native one (Fig. 1*d*). The domains that are significant for positive patches formation usually not involved in aminoacylation reaction and demonstrate low SI values. Some of them interact with tRNA, whereas function of others hasn't been detected yet.

The Cp1 domain of IleRS that may be considered as a "crucial" for characteristic positive patch formation (see Fig. 1*b*) is associated with additional proofreading activity of the enzyme and contains a distinct active site where misactivated aminoacyl-adenylate or misaminoacylated tRNA are hydrolyzed (Silvian *et al.*, 1999).

In PheRS isolated from *Thermus thermophilus*, the coiled-coil domain of α-subunit (CC*) is characterized by SI close to 0 (Mosyak *et al*., 1995). Two positive patches on RES, corresponding to two cognate tRNA interacting with PheRS becomes distorted and vanishingly small when charged residues of coiled-coil are switched-off. Functions of B2 domain that also plays a significant (albeit less pronounced) role in EP formation are not immediately evident from structures of the various functional complexes (Goldgur *et al*., 1997). A similar observation hold true for other aaRSs. Thus, domains that contribute significantly to the "blue space" formation very often involved in alternative activities of the

aaRSs. It is notable that structural domains of aaRSs playing a 'crucial' role in tRNA-protein recognition at long distances have a relatively low concentration of conserved charged residues. Therefore, NCS arranged in these domains can be treated as positive electrostatic determinant favoring the attraction and navigation of tRNA to its binding area.

It is possible to speculate that aaRSs family has acquired these domains at later stages of evolution.

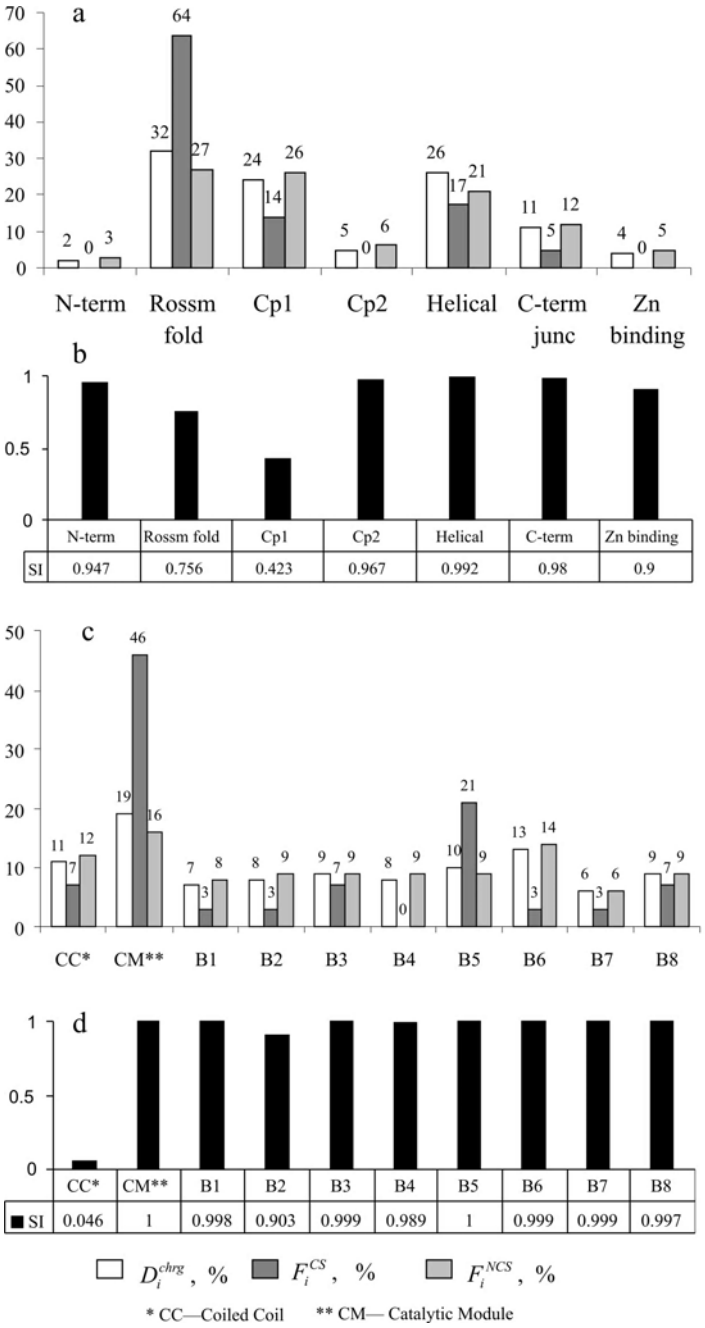


Figure 1. Distribution of charged residues and SI values in IleRS (*a*, *b*) and PheRS (*c*, *d*) domains.

REFERENCES

Berg O.G., von Hippel P.H. (1985) Diffusion-controlled macromolecular interactions. *Ann Rev. Biophys Biophys Chem*., **14**, 131–160.

Blomberg N., Gabdoulline R.R., Nilges M., Wade R.C. (1999) Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Proteins*, **37**, 379–387.

Goldgur Y., Mosyak L., Reshetnikova L., Ankilova V., Khodyreva S., Lavrik O., Safro M. (1997) The crystal structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus* complexed with cognate $tRNA^{Phe}$. *Structure*, **5**, 59–68.

Mosyak L., Reshetnikova L., Goldgur Y., Delarue M., Safro M. (1995) Structure of phenylalanyl-tRNA synthetase from *Thermus thermophilus*. *Nat. Struct. Biol.*, **2**, 537–547.

Rocchia W., Sridharan S., Nicholls A., Alexov E., Chiabreba A., Honig B. (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. *J. Comput. Chem.*, **23**, 128–137.

Silvian L., Wang J., Steitz A. (1999) Insights into editing from an Ile-tRNA synthetase structure with $tRNA^{Ile}$ and mupirocin. *Science*, **285**, 1074–1077.

Tworowski D., Safro M. (2003) The long-range electrostatic interactions control tRNA–aminoacyl-tRNA synthetase complex formation. *Prot. Scie*., **12**, 1247–1251.

Tworowski D., Feldman A., Safro M. (2005) Electrostatic potential of aminoacyl-tRNA synthetase navigates tRNA on its pathway to the binding site. *J. Mol. Biol*., **350**, 866–882.

MOLECULAR DYNAMICS AND DESIGN OF TRANSMEMBRANE ION CHANNELS

Shaitan K.V., Tereshkina K.B., Levtsova O.V.
Department of Bioengineering, Faculty of Biology, M.V. Lomonosov Moscow State University, 119992, Moscow, Russia
* Corresponding author: e-mail: shaitan@moldyn.org

Key words: steered molecular dynamics (SMD), membrane receptors, ion channels, ion migration

SUMMARY

In this work functioning of ion channels is studied using steered molecular dynamics method. Gramicidin A channel along with glycine and acetylcholine receptor channels are considered in terms of ion migration and conformation dynamics.

INTRODUCTION

The progress in theory and computer sciences gives us new opportunities for development of biomolecular design. For these purposes, MD simulation and its different variations are used. Below, we consider functions of transmembrane ion channels. These channels are usually formed as an association of peptide structures. Ion selectivity of a channel depends on the amino acid sequence. Varying this sequence, the desired selectivity can be achieved. For example, it is important for design of biosensors. The same methods can be used to for search of targets and appropriate ligands as well.

METHODS AND ALGORITHMS

The technique is based on numerical solution of classical Newtonian equations with respect to a many-atom system. It is well known from mechanics that the evolution of a system and all its properties is totally predefined by the interaction potentials and initial conditions. Thus, solving these equations of motion for a many-atom system we can obtain its evolution and properties. At present MD simulations are applied to many issues in the areas from biology up to astrophysics. In most cases the simulation results are rather reliable and can be regarded as derived from a computational experiment. There are two variants of MD simulations: the one is equilibrium MD, and the other is Steered MD. In SMD we apply definite external forces to certain particles or apply extra boundary conditions, depending on the problem. For study of large molecular systems the SMD approach is preferable. This is due to the problem of ergodicity and impossibility to reach equilibrium state *in silico* in a reasonable time period (Shaitan, Tereshkina, 2005).

RESULTS AND DISCUSSION

Let's consider one of the simplest channels formed by two molecules of gramicidin A. Here we can see the channel itself embedded into a phospholipidic bilayer (Fig. 1*a*), and

separately the structure of the channel in two projections (Fig. 1*b*, *c*). The balls are carbonyl oxygen atoms of peptide groups. These groups form layers in the channel. There is an excess of electronic density on the carbonyl groups which impacts the kinetics of cation transfer through the channel.

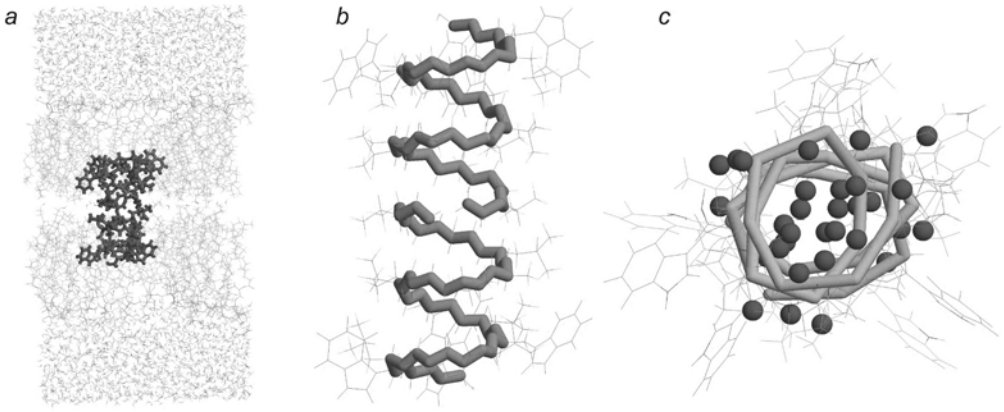


Figure 1. Structure of the system (*a*) in two projections (*b*, *c*).

Sodium ion transfer along the channel is represented on Fig. 2. The ion hydrated by six water molecules enters the channel and looses four water molecules. Then the ion hydrated by two molecules of water moves through the channel. For a better visual perception the membrane and water atoms are not depicted. The ion movement takes place under the action of an external force or electric potential.

In the areas where the cation comes across the excess negative charge density of carbonyl groups, we can observe that its motion slows down (Fig. 2). In other words, the peculiarities of the channel interior are rather important, and this can be a matter of molecular design.

Then we consider an example of anion channel, formed by TM2 subunits of glycine receptor (Yushmanov *et al*., 2003). The receptor plays an important role in the functioning of nervous system. The channel is formed by five transmembrane alpha-helices (Fig. 3). As in the previous case there is a region of excess charge, but now it is positive charge from arginine residues that play definite role in chlorine anion transfer. Mutations in the channel part of the receptor affect the conductivity of the channel, and can even transform the channel from an anionic to cationic one.

On Fig. 4 we can see *in silico* reconstruction of the channel part of the receptor that we have studied. The channel is represented by a funnel formed by 5 alpha-helixes. The funnel entrance for the ion is quite wide; the funnel outlet is rather narrow. For stabilization of the system we used an alkane rim which played a role of bandage. On Fig. 4*b* you can see a pentameric structure of the channel, view from the outlet side.

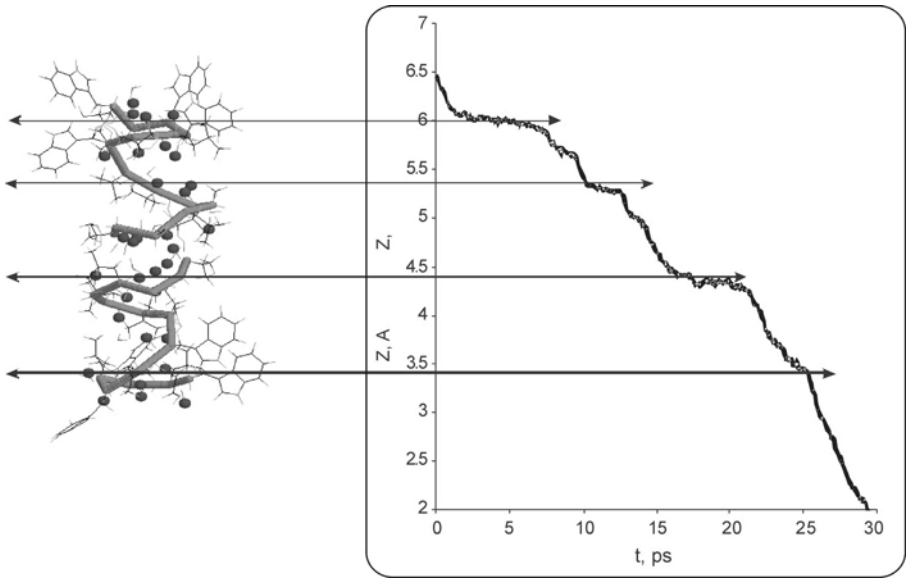


Figure 2. Kinetics of the sodium ion transfer.

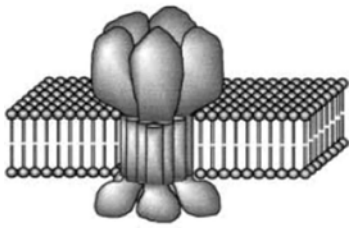


Figure 3. General view of the glycine receptor in membrane.

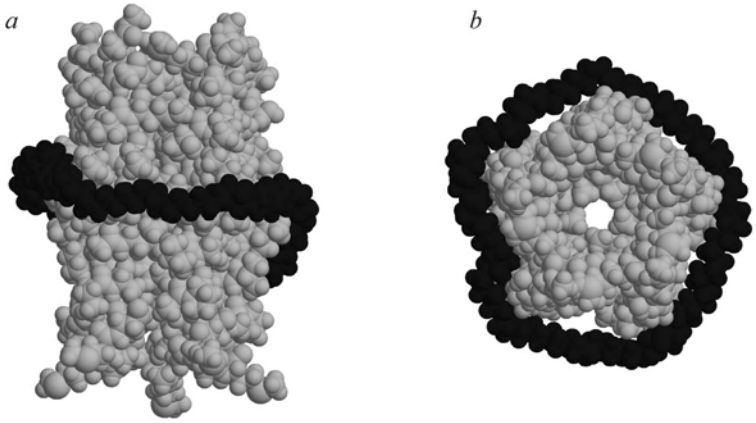


Figure 4. Model of the glycine receptor's channel. Side view (*a*). Top view (*b*).

The dynamics of chlorine anion transfer through the channel can be divided into three phases (Fig. 5). At the beginning we can observe a rather slow entering of the anion into the channel through the first belt of arginine residues. Then a rapid transfer up to the lower belt takes place. Then a rather slow phase occurs, when the anion overcomes this barrier and escapes the channel.

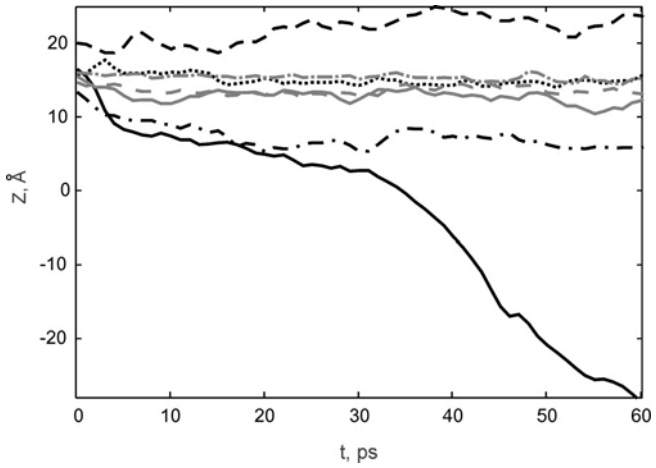


Figure 5. Kinetics of the chlorine ion (solid line) and the hydration shell water molecules (other liner) transfer through the channel of the glycine receptor.

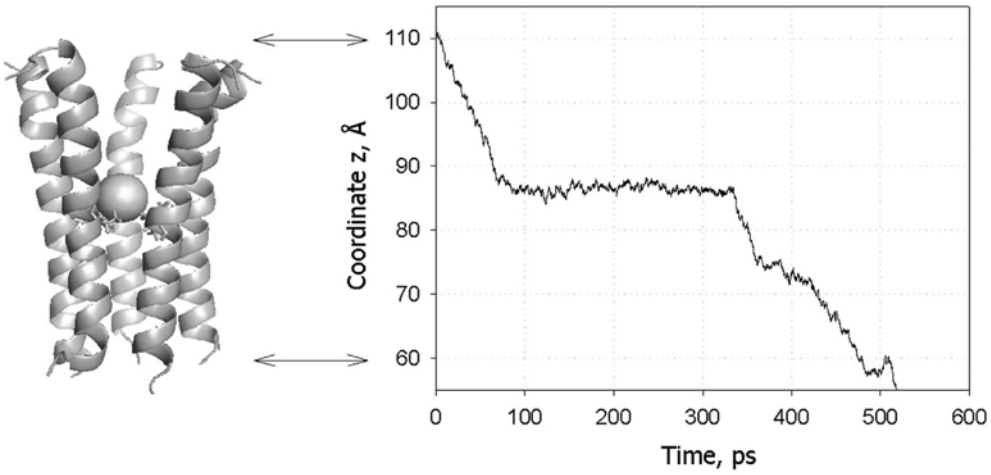


Figure 6. Model of the acetylcholine receptor's channel and kinetics of the sodium ion transfer.

The kinetics of the process is shown on Fig. 5 (the position of the chloride ion is shown by a solid black line).

On Fig. 6 we can see the migration process of a sodium ion through the nicotine acetylcholine channel, a channel of the same family as glycine one (Miyazawa *et al*., 2003). It seems that the interior of ion channels is constructed in such a way that charged side chains of amino acid residues tend to form some kind of gates responsible for the channels' selectivity.

CONCLUSION

Computer design of charge interior of the channels by means of point mutations seems to be very useful for design of channels with predefined properties. SMD simulations is the most efficient and economical way to find some possible new products in this area.

ACKNOWLEDGEMENTS

The work was supported by RF Federal Agency on Science and Innovation, Russian Foundation for Basic Research (project No. 04-04-49645) and US CRDF.

REFERENCES

Miyazawa A., Fujiyoshi Y., Unwin N. (2003) Structure and gating mechanism of the acetylcholine receptor pore. *Nature*, **423**, 949–955.

Shaitan K.V., Tereshkina K.B. (2005) Molecular Dynamics of Small Peptides Using Ergodic Trajectories In Kolchanov N., Hofestaedt R. (eds.) *Bioinformatics of Genome Regulation and Structure II*, Springer Science+Business Media. pp. 271–284.

Yushmanov V.E. *et al*. (2003) NMR Structure and backbone dynamics of the extended second transmembrane domain of the human neuronal glycine receptor 1 subunit. *Biochemistry*, **42**, 3989–3995.

PEPTIDE DYNAMICS AT WATER-MEMBRANE INTERFACE

Shaytan A.K.*, Khokhlov A.R., Ivanov V.A.
Moscow State University, Moscow, Russia
* Corresponding author: e-mail: shaytan@polly.phys.msu.ru

Key words: amino acids, solvation free energy, adsorption, molecular dynamics

SUMMARY

Motivation: Elaboration of computer aided molecular modeling techniques for prediction of solute behaviour in biphasic (water/hydrophobic medium) systems. The main focus is on biological systems and especially drug design.

Results: molecular dynamics simulations for all main types of amino acid residues at water/vacuum and water/hexane interface were performed. We compute distribution and orientation histograms for amino acid residues to analyze their behaviour in the biphasic system.

INTRODUCTION

It is well known that one of the most important structure formation components in biological systems is biomembranes. They in fact form surfaces that can be treated as interfaces between polar aqueous solutions and non-polar environment represented by phospholipidic side chains. It is hard to overestimate the role of polarity and non-polarity of phases on the distribution of substances and formation of different structures inside biological systems. Of special interest are the interaction properties of proteins with solvents and interfaces (Okhapkin *et al*., 2005, 2006) because these interactions contribute to the process of protein folding and govern its functional activity. That's why our idea was to study the behaviour of amino acid residues, as monomer units of proteins, in an interfacial system.

OBJECTIVE

By means of molecular dynamics simulations the properties of amino acid residues in a system containing liquid-liquid interface between water and non-polar phase are studied (Fig. 1). The non-polar phase is modeled either by hexane or vacuum media. We intend to determine free energy profiles, free energy of partitioning between two phases and free energy of adsorption at interface. MD simulations can also reveal conformational changes of residue structure in different environments.

SIMULATION SECTION

The system consisted of two lamellae (one of water and one of hexane or vacuum) (Fig. 1), placed in a simulation with periodic boundary conditions in all three directions. In

fact there are two interfaces because of the periodicity. The x and y box sizes are 40 A. The height of each lamella is approximately 20–36 A, but it varies slightly because of the Berendsen barostat algorithm applied along z-axis maintaining pressure of 1 bar. There are 1070 water molecules and 147 hexane molecules in hexane lamella. Hexane phase was initially arranged in 3 layers of hexane molecules, 7*7 = 49 molecules in each layer. We make simulations with collision thermostat, maintaining system temperature at 300K. Integration step of 1 fs was chosen. Trajectories up to several nanoseconds were collected.

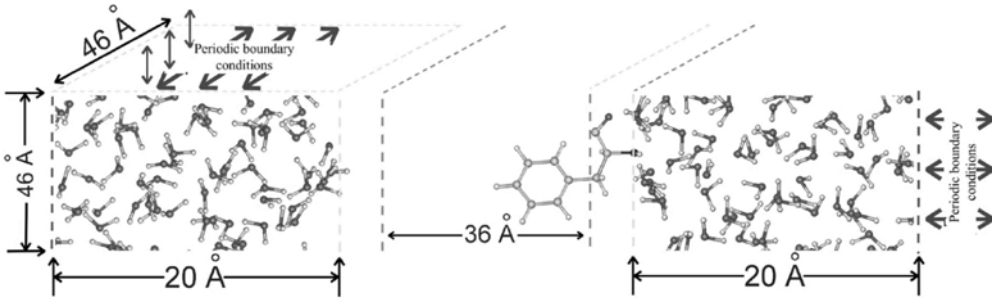


Figure 1. Structure of water/vacuum interface system with amino acid residue.

We used AMBER99 forcefield and TIP3P water model with unconstrained internal degrees of freedom. The amino acid residue, initially settled at the interface, was used in the form as if they were built into protein sequence. We did not do any terminal blocking. In fact they cannot exist in reality, but they reflect the properties of protein monomer units.

As a first step towards modeling the above described system was modeling it without hexane phase, but using a rigid wall potential to prevent water molecules evaporate from the surface. From the simulations we got residue trajectories, derived probability histograms for residue and for its parts. We built up orientation patterns of residues.

RESULTS AND DISCUSSION

Let us consider at first the results of water-vacuum simulations. We performed them for all main types of residue side chains. Generally speaking, we can regard side chains as hydrophobic or hydrophilic to some extent.

For purely hydrophobic side chains the residues tended to stay near the interface. We consider an example of phenylalanine (Fig. 2). Probability histograms for mass center of backbone and side chain are depicted. Two vertical lines bound the region of repulsive potential. We see that PHE adsorbs on the interface. And it is clear that it has a certain conformational and rotational asymmetry. The backbone being a hydrophilic part tends to settle itself deeper in the water layer than the side chain.

The similar behaviour was observed in systems with other residues. But as for residues with strongly hydrophilic side chains, they tended to desorb from the interface into bulk water phase.

Let us now pass to a water-hexane system.

On the Fig. 3 we see the density profiles of hexane and water phases. The interface region is about 5 A wide. On Fig 4 we observe the probability histogram for this system. The residue adsorbs on the interface and again has dominant orientation direction. But the distance between peaks of two curves is about 1.2 A instead of 2 A in the water-vacuum case.

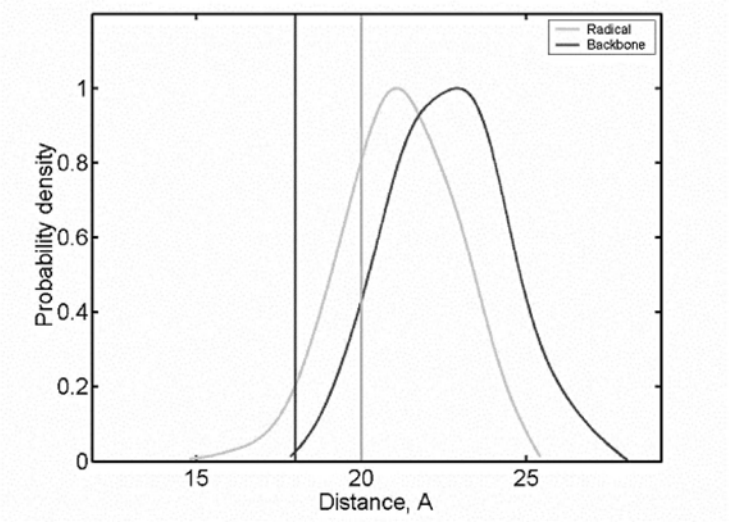


Figure 2. Distribution functions for backbone and side chain of PHE in water/vacuum system simulations.

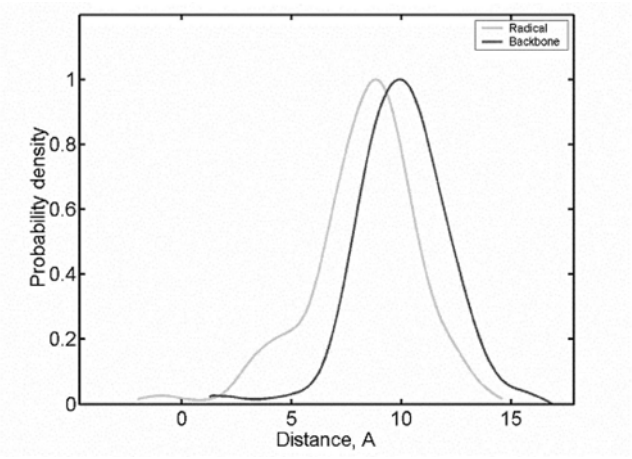


Figure 3. Distribution functions for backbone and side chain of PHE in water/hexane system simulations.

Analogously other types of residues were studied.

Analyzing histograms for different types of residues both in water/vacuum and water/hexane studies we find that behaviour of amino acid residues at water/vacuum and water hexane interfaces is in general similar.

We propose that it is possible to use water/vacuum systems with collision thermostat as rough model for water/hydrophobic medium interface systems.

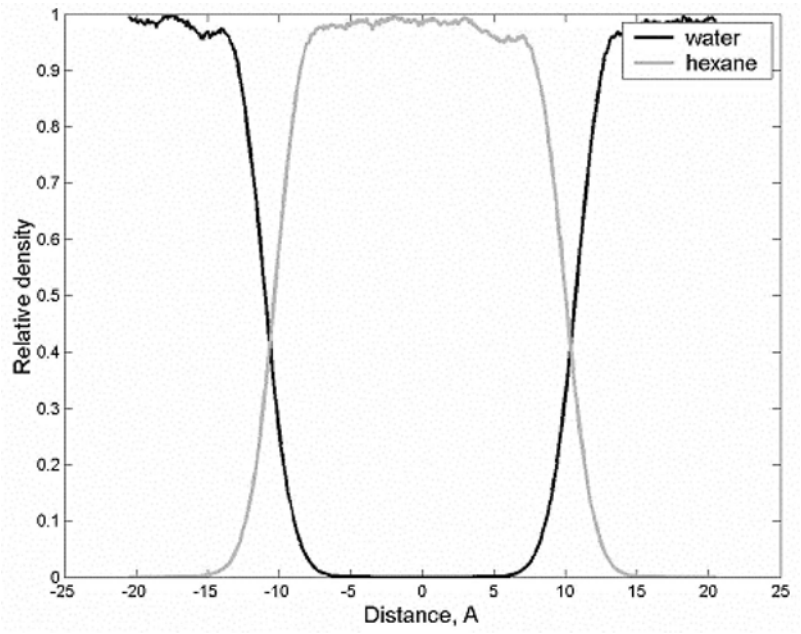


Figure 4. Relative density profiles for water and hexane in biphasic system.

REFERENCES

Okhapkin I.M., Askadskii A.A., Markov V.A., Makhaeva E.E., Khokhlov A.R. (2006) Two-dimensional classification of amphiphilic monomers based on interfacial and partitioning properties. 2. Amino acids and amino acid residues. *Colloid and Polymer Sci*., **284**(6), 575–585.

Okhapkin I.M., Makhaeva E.E., Khokhlov A.R. (2005) Two-dimensional classification of amphiphilic monomers based on interfacial and partitioning properties. 1. Monomers of synthetic water-soluble polymers. *Colloid and Polymer Sci*., **284(**2), 117–123.

CONFORMATION PROPERTIES OF SHORT OLIGOPEPTIDES AND PREDICTION OF PROTEIN CHAIN CONFORMATION

Vlasov P.K.*[*], *Esipova N.G.*, *Tumanyan V.G.
Engelhardt Institute of Molecular Biology, RAS, Moscow, Russia
[*] Corresponding author: e-mail: vlasov@imb.ac.ru

Key words: protein secondary structure, oligopeptide, conformation, prediction, left-handed helix of poly-L-proline II type

SUMMARY

Motivation: Modern methods of protein secondary structure prediction, based entirely on protein sequences, have a very good results for "typical" α- and β-structures. But there are no accurate prediction methods for other types of the protein chain conformation, firstly for polyproline II left-helical conformation (PPII). However, PPII conformation has a very important biological role. New approaches of protein chain conformation annotation are required for adequate fold recognition and modeling.

Results: The different conformation type fragments in the globular proteins of the protein databank (PDB) were analyzed. We revealed the interrelation between sequence and structure even for very short oligopeptides. It was found the tetrapeptides with a good preference of distinct types of secondary structure. It is the first method for structure annotation with a relatively high accuracy level (~60 %) for PPII conformation prediction.

Availability: WEB-server containing tetrapetide structure properties databank and search tool: http://strand.imb.ac.ru/consol/index.html. Protein chain conformation prediction method: http://strand.imb.ac.ru/ssp/index.html.

INTRODUCTION

Modern methods of protein chain conformation analysis and prediction aimed to long-range fragments of a protein chain, big domains and regular structure segments. There are now available many methods with an accuracy level of ~80 % for α-helices and β-sheets (King *et al*., 2000). However, new approaches are required that can reveals interrelation between sequence and secondary structure (Koehl, Levitt, 1999; Rost, 2001; Aydin *et al*., 2006; Lee *et al*., 2006). Moreover, there are many protein chain segments with a non-regular structure (β-turns), and there is a specific conformation of poly-L-proline-II type having small size of typical element (three-four residues) and absence of inner hydrogen bonds. It is clear now that left-helical conformation (PPII) plays an important structural and biological role (Blanch *et al*., 2000; Vlasov *et al*., 2001; Rath *et al*., 2005). So it seems reasonable to analyze conformational properties of short chain fragments and design a prediction algorithm for non-regular conformations. We try in our work to summarized conformation properties of short oligopeptides (di-, tri- and tetrapeptides) of globular proteins from PDB to select oligopeptides with stable (predictable) conformations.

METHODS AND ALGORITHMS

The relative content for different types of conformations in globular proteins was estimated and a big ratio of left-handed helical segments was shown. We analyzed conformation properties of oligopeptides, and shows existence of tetrapeptides with good preference of distinct structure type: α-, β- and left-handed helix.

We propose a new method for secondary structure prediction based on oligopeptide conformation properties. This method decomposes sequence under analysis into overlapping tetrapeptides, and each residue is present in four tetrapeptide fragments. For example, the D residue at the sequence ARNDCEV occurs in ARND, RNDC, NDCE and DCEV tetrapeptides. We estimate the probability with which this residue (D), from the above sample sequence, forms a particular structure, using a simple additive scheme:

$$P(struct)_D = ¼(P_{ARND} + P_{RNDC} + P_{NDCE} + P_{DCEV}),$$

there $P(struct)_{ARND}$, $P(struct)_{RNDC,}$ $P(struct)_{NDCE}$ and $P(struct)_{DCEV}$ are frequencies with which residue (D) forms a particular structure in corresponding tetrapeptides in all globular proteins of PDB.

IMPLEMENTATIONS AND RESULTS

The special database ConSOL (Conformation Statistic of Oligopeptides) was created and corresponding WEB-service was design. This relational database containing information about tetrapeptide structure preferences: occurrence in globular protein chains, frequencies of three main structure types (α-, β- and left helix) separately for every position (1st, 2nd, 3rd, and 4th residue) and common oligopeptide structure description (such as “all-α” or “mixed-β-and-left-helical”).

Using the test subset of distantly related proteins whose secondary structure was determined with a high accuracy, we estimates average values of the prediction accuracy for main three types of secondary structures:

- α-helices: ~ 72 %,
- β-sheets: ~ 70 %,
- PPII conformation: ~65 %.

ConSOL server and a new method of secondary structure prediction are described in (Vlasov *et al.*, 2005).

DISCUSSION

We suggest that even short sequence segments (tetrapeptides) provide sufficient information for predicting protein chain conformation. Our aim was to investigate structural preference of short oligopeptides, and in contrast to common prediction schemes, we use the observed conformational preferences without additional rules or parameters. In our approach, at the first time good prediction for poly-L-proline-II structure was obtained. Although empirical rules may improve formal prediction results, it would obscure the described sequence-structural biases and, therefore, we have not done this.

The ConSOL database will be useful to protein design and analysis of peptide chain conformations. It is possible now to compare conformation preferences of short peptide fragments and select oligopetides of specific conformation type. Analysis of most

structure-stable oligopeptides may contribute to understanding how amino acid substitutions may affect a protein chain local conformation.

ACKNOWLEDGEMENTS

The authors are grateful to F. Kondrashov for critical analysis of the results. This work was supported by the grant from the Russian Foundation for Basic Research (No. 03-04-49017) and the grant on Molecular and Cellular Biology RAS (Program No. 10).

REFERENCES

Aydin Z., Altunbasak Y., Borodovsky M. (2006) Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics*, **7**(1), 178.

Blanch E.W., Morozova-Roche L.A., Cochran D.A.E., Doig A.J., Hecht L., Barron L.D. (2000) Is polyproline II helix the killer conformation? A Raman optical activity study of the amyloidogenic prefibrillar intermediate of human lysozyme. *J. Mol. Biol.*, **301**, 553–563.

King R.D., Ouali M., Strong A.T., Aly A., Elmagharby A., Kantardzic M., Page D. (2000) Is it better to combine predictions? *Protein Engineering*, **13**(1), 15–19.

Koehl P., Levitt M. (1999) A brighter future for protein structure prediction. *Nature Struct. Biol.*, **6**, 108–111.

Lee S., Lee B.C., Kim D. (2006) Prediction of protein secondary structure content using amino acid composition and evolutionary information. *Proteins*, **62**(4), 1107–1114.

Rath A., Davidson A.R., Deber C.M. (2005) The structure of "unstructured" regions in peptides and proteins: role of the polyproline II helix in protein folding and recognition. *Biopolymers*, **80**(2/3), 179–185.

Rost B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**(2/3), 204–218.

Vlasov P.K., Kilosanidze G.T., Ukrainskii D.L., Kuzmin A.V., Tumanyan V.G., Esipova N.G. (2001) Left-handed Conformation of Poly-L-proline-II Type in Globular Proteins. *Sequence Specificity, Biophysics*, **46**(3), 573–576.

Vlasov P.K., Vlasova A.V., Tumanyan V.G., Esipova N.G. (2005) A tetrapeptide-based method for polyproline II-type secondary structure prediction. *Proteins: Structure, Function and Bioinformatics*, **61**(4), 763–768.

AMINO ACID PREFERENCES AT THE N-TERMINAL PART OF EUKARYOTIC PROTEINS CORRELATING WITH A SPECIFIC CONTEXTUAL ORGANIZATION OF TRANSLATION INITIATION SIGNAL

Volkova O.A., Kochetov A.V.*
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
* Corresponding author: e-mail: ak@bionet.nsc.ru

Key words: translation initiation, mRNA, N-end of protein, amino acids composition

SUMMARY

Motivation: It was shown that the nucleotide sequence at the CDS beginning as well as N-terminal amino acids could influence the recognition of translation start site. However, interrelationship between these features and mRNA translation initiation efficiency was not investigated in detail.

Results: Statistical deviations in amino acid frequencies at N-terminal positions 2–4 of proteins were analyzed in *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Homo sapiens* genes. It was found that the most frequent amino acid combinations in these positions were species-specific (Ala-Ser-Ser for *Arabidopsis thaliana*, Ser-Ser-Asn for *Drosophila melanogaster*, and Ala-Ala-Ala for *Homo sapiens*). Note that the second position of the protein sequence was most frequently occupied by amino acids with codons starting from GN but not GU. Statistically significant deviations in N-terminal amino acid frequencies are likely to correlate with their influence on both protein stability and translation start site recognition.

INTRODUCTION

Contextual organization of the 5′-terminal part of the protein coding sequence (CDS) may result from several factors including the usage of a protein-specific amino acids, preferable usage of optimal codons to provide mRNA with a high translation elongation rate, and a specific organization of the translation initiation signal. It was shown that the positional nucleotide frequencies in this region are highly biased. The most biased CDS position is the 2nd triplet of CDS (Berezovsky *et al*., 1999; Sawant *et al*., 2001; Niimura *et al*., 2003). It was demonstrated that the most frequent amino acid at this position were alanine for arabidopsis and human proteins and serine for drosophila (Sawant *et al*., 2001; Niimura *et al*., 2003). The most frequent amino acids at positions 2–4 of proteins in plants were reported to be Ala – Ser – Ser (Sawant *et al*., 2001) and, in human proteins, the second position was most frequently occupied by threonine (Berezovsky *et al*., 1999). It is assumed that these amino acids increase protein stability but this hypothesis has not been verified.

MATERIALS AND METHODS

Dataset

The cDNAs of *Arabidopsis thaliana*, *Drosophila melanogaster*, and *Homo sapiens* genes were extracted from the EMBL database (04.02.2005 release). The redundant nucleotide sequences (with homology more than 95 %) were excluded with the aid of the CleanUp program (Grillo *et al*., 1996). The resulting sets included 13768 cDNAs of *Arabidopsis thaliana*, 2005 of *Drosophila melanogaster*, and 24414 of *Homo sapiens*. The average codon frequencies (Exp^{CUTG}, frequency per 1000 codons) were obtained from http://www.kazusa.or.jp/codon/.

Statistical analysis

The expected (Exp) codon frequencies were calculated as $Exp = 0.001*N*Exp^{CUTG}$ (N, number of sequences in a sample). The observed (Obs) codon frequencies were calculated as Obs = Ai/0.001*N, (where Ai is the observed number of codons of type A at position *i*, N is the number of nucleotide sequences in a set), the Obs and Exp frequencies for amino acids residues were calculated as the total sum of the Obs and Exp frequencies of the corresponding synonymous codons. For each amino acid residue at a given position, the deviation of the Obs values from the Exp values was estimated by χ^2 criterion according to: $\chi^2 = (Obs-Exp)^2/Exp$. For each amino acid residue, the χ^2 value was estimated separately with one degree of freedom. The sum of all 20 (61) χ^2 values for each residue (codon) at a given position gave the total deviation for the given position with 19 (60) degrees of freedom. In this work, we used the following numeration: the start AUG codon is numbered +1; the first nucleotide of CDS (A in AUG codon), +1.

RESULTS AND DISCUSSION

We evaluated the deviations of amino acid frequencies from the expected values. It was found that the amino acid content at positions 2–4 was significantly biased (Table 1). Note that the amino acid frequencies were more biased at pos. 2 than at pos. 3, and were more biased at pos. 3 than at pos. 4.

Table 1. The χ^2 values at positions 2–4 of proteins

Position	Second	Third	Fourth
Arabidopsis thaliana	9409.08	1167.89	503.97
Drosophila melanogaster	467.14	85.81	59
Homo sapiens	9661.52	1329.86	740.67

The ratio of observed to expected amino acid frequencies and the significance of deviations from the expected values were calculated (Table 2).

There are some preferences common for all organisms analyzed. For example, amino acid preferences at pos. 2 is Ala> Ser> Glu*>Gly*> Asp (* - except for *D. melanogaster*). It can be noted that the most frequent amino acids corresponded to the codons starting from G (Ala, Glu, Gly, Asp). This might result from the functional significance of guanine at pos. +4 of CDS (*i.e.*, at the 1st nucleotide of a second codon) (Kozak, 1997). The most overrepresented second codon is GCG: probably, this triplet provides more efficient AUG recognition compared with other GNN combinations. However, the codons of valine also start from G, but this amino acid is underrepresented at the 2nd protein position (Table 2). It is likely that the valine-encoding codons are unfavorable for the translation initiation rate. It had previously been shown in mammalian *in vitro* translation systems that the positive influence of guanine at pos. +4 was eliminated if position +5 was occupied by uridine (Kozak, 1997). Valine codons corresponded to this situation (GUN) and this could be the reason for valine underrepresentation. Our results suggest that

$G^{+4}U^{+5}$ combination is unfavorable *in vivo* not only in mammalian (Kozak, 1997) but also in drosophila and arabidopsis mRNAs.

Table 2. The ratio of observed to expected amino acid frequencies at pos. 2, 3, 4 of CDS of arabidopsis, human, and drosophila genes

Obs/Exp	*Arabidopsis thaliana*			*Drosophila melanogaster*			*Homo sapiens*		
Position	2	3	4	2	3	4	2	3	4
Lys	*0.79*[1] *	**1.10**[2]	1.06	0.86	1.12	**1.26**[3]	*0.86*[1]	*0.89*[1]	0.99
Thr	*0.79*[1]	**1.48**[1]	**1.25**[1]	1.14	**1.31**[1]	1.14	*0.90*[1]	**1.15**[1]	1.00
Asn	*0.70*[1]	**1.09**[4]	1.03	*0.79*[5]	1.06	**1.42**[1]	*0.91*[3]	*0.81*[1]	*0.88*[1]
Met	0.96	**1.17**[2]	0.93	0.99	0.86	1.14	*0.67*[1]	*0.79*[1]	*0.76*[1]
Ile	*0.44*[1]	*0.83*[1]	1.01	*0.37*[1]	0.82	1.06	*0.40*[1]	*0.57*[1]	*0.72*[1]
Glu	**1.65**[1]	1.02	0.97	1.15	0.88	0.99	**1.32**[1]	1.02	*0.94*[4]
Asp	**1.25**[1]	*0.84*[1]	*0.71*[1]	**1.41**[1]	**1.25**[3]	0.86	**1.19**[1]	0.99	*0.73*[2]
Gly	**1.48**[1]	*0.92*[4]	*0.84*[1]	0.94	*0.70*[1]	*0.67*[1]	**1.21**[1]	**1.10**[1]	1.01
Ala	**3.73**[1]	**1.08**[4]	0.95	**1.92**[1]	1.16	0.95	**3.02**[1]	**1.50**[1]	**1.28**[1]
Val	*0.80*[1]	*0.69*[1]	*0.90*[2]	*0.62*[1]	0.72	0.98	*0.71*[1]	*0.81*[1]	*0.88*[1]
Gln	*0.59*[1]	*0.77*[1]	**1.05**[1]	*0.75*[3]	*0.75*[3]	0.91	*0.57*[1]	*0.91*[2]	*0.92*[3]
His	*0.31*[1]	*0.68*[1]	*0.82*[2]	*0.68*[4]	0.94	0.73	*0.41*[1]	*0.75*[1]	*0.78*[1]
Arg	*0.56*[1]	**1.21**[1]	**1.06**	*0.69*[2]	1.11	0.96	*0.77*[1]	**1.21**[1]	**1.24**[1]
Pro	*0.50*[1]	*0.71*[1]	*0.88*[2]	0.97	0.86	1.04	*0.92*[2]	**1.15**[1]	**1.25**[1]
Tyr	*0.32*[1]	*0.55*[1]	*0.71*[1]	0.81	0.81	*0.67*[4]	*0.36*[1]	*0.51*[1]	*0.68*[1]
Trp	*0.35*[1]	*0.63*[1]	*0.62*[1]	0.66	0.86	1.16	*0.95*	1.11	**1.14**[4]
Cys	*0.35*[1]	0.89	0.93	*0.40*[1]	0.99	0.91	*0.44*[1]	*0.77*[1]	*0.86*[1]
Ser	**1.20**[1]	**1.67**[1]	**1.47**[1]	**2.06**[1]	**1.35**[1]	0.99	**1.39**[1]	**1.27**[1]	**1.18**[1]
Leu	*0.36*[1]	*0.85*[1]	1.03	*0.66*[1]	0.93	1.03	*0.56*[1]	0.96	**1.07**[1]
Phe	*0.43*[1]	*0.82*[1]	0.98	*0.67*[2]	1.15	1.19	*0.57*[1]	*0.74*[1]	0.96

* [1], [2], [3], [4], [5] - significance levels $P < 0.05$, 0.025, 0.01, 0.005, 0.001, respectively.
Significantly overrepresented values are **bold-faced**, significantly underrepresented, *italicized*.

Valine underrepresentation at pos. 2 of eukaryotic proteins may also in part result from its hydrophobicity. As can be seen, hydrophobic amino acids Leu, Ile, Phe, Trp, Pro, Val are mainly underrepresented at N-terminal positions. Interestingly, serine is also overrepresented at pos. 2 and its codons (especially UCG) are overrepresented in all organisms analyzed. This is very likely to be an exception, since these codons start from nucleotides other than guanine. It is possible that UCG also increase AUG recognition because some specific features of this triplet may abrogate the negative influence of uridine at position +4. It may also be assumed that the combination Met – Ser at N-end is important for the functional activity of certain protein classes, hence a suboptimal AUG context. Generally, amino acids overrepresented at the 2nd position (Ser, Glu, Gly, Asp) are hydrophylic (Sweet, Eisenberg, 1983): note that both negatively charged amino acids (Glu, Asp) appeared in this group. It was also shown that the presence of Ala, Ser, Thr, Gly, and Pro at the N-end of proteins increases cytoplasmic stability (Tobias *et al*., 1991; Varshavsky, 1996; Sawant *et al*., 2001).

The 3rd position of proteins is characterized by preferable usage of Ser, Thr and Arg* (* except for *D. melanogaster*). The 4th position is characterized by lesser deviations in amino acid frequencies from the expected values and no common pattern for the organisms assayed was found. The most overrepresented are Ser and Thr (arabidopsis), Asn and Lys (drosophila), Ala and Pro (human).

ACKNOWLEDGEMENTS

We thank F.A. Kolpakov for program developing and V.A. Ivanisenko for helpful discussion. This work was supported by the Russian Foundation for Basic Research (No. 05-04-48207) and RAS program (Dynamics of Plant, Animal and Human Gene Pools). We thank SD RAS Complex Integration Program (N5.3), and Ministry of Industry, Sciences and Technologies of Russian Federation (2275.2003.4) for partial support.

REFERENCES

Berezovsky I.N., Kilosanidse G.T., Tumanyan V.G., Kisselev L.L. (1999) Amino acid composition of protein termini are biased in different manners. *Protein Engineering*, **12**, 23–30.

Grillo G., Attimonelli M., Liuni S., Pesole G. (1996) CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases. *Comp. Appl. Biosci*., **12**, 1–8.

Kozak M. (1997) Recognition of AUG and alternative initiator codons is augmented by G in position +4 but is not generally affected by the nucleotides in positions +5 and +6. *EMBO J.*, **16**, 2482–2492.

Niimura Y., Terabe M., Gojobori T., Miura K. (2003) Comparative analysis of the base biases at the gene terminal portions in seven eukaryote genomes. *Nucl. Acids Res.*, **31**, 5195–5201.

Sawant S.V., Kiran K., Singh P.K., Tuli R. (2001) Sequence Architecture Downstream of the Initiator Codon Enhances Gene Expression and Protein Stability in Plants. *Plant. Physiol*., **126**, 1630–1636.

Sweet R.M., Eisenberg D. (1983) Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.*, **171**, 479–488.

Tobias J.W., Shrader T.E., Rocap G., Varshavsky A. (1991) The N-end rule in bacteria. *Science*, **254**, 1374–1377.

Varshavsky A. (1996) The N-end rule: Functions, mysteries, uses. *Proc. Natl. Acad. Sci. USA*, **93**, 12142–12149.

PROBING DIMERIZATION OF TRANSMEMBRANE PEPTIDES *VIA* MOLECULAR DYNAMICS IN EXPLICIT BILAYERS

Volynsky P.E.**, Vereshaga Ya.A., Nolde D.E., Efremov R.G.
M.M. Shemyakin and Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, RAS, Moscow, Russia
* Corresponding author: e-mail: pashuk@nmr.ru

Key words: membrane proteins, oligomerization, helix-helix interactions, protein kinase

SUMMARY

Motivation: Often, dimerization (oligomerization) of membrane proteins (MPs) is crucially important for their functioning. Investigation of such processes by modern experimental methods is hampered by a number of technical difficulties. An alternative approach lies in elaboration of molecular modeling (or, *in silico*) techniques.

Results: A computational approach to study the dimerization of α-helical MPs is proposed. The method is based on molecular dynamics (MD) of peptides in model lipid bilayers. It was tested via exploration of the structure of a dimer formed by TM fragments of pro-apoptotic protein BNIP3. The results of simulations are in a good agreement with the experimental data. Moreover, MD simulations allow assessment of some effect not detectible by the experimental techniques.

INTRODUCTION

MPs play a key role in cell life being involved in signaling, ion conductance across the membrane, cell communications, membrane transport, and so on. Unique properties of these proteins are determined mainly by the spatial structure and dynamic behavior of their transmembrane (TM) domains. α-Helix represents one of the most frequently occurring structural motifs in TM protein fragments. Many MPs (e.g., ion channels, receptors, etc) consist of a bundle of helices. Moreover, in many cases dimerization (oligomerization) of MPs is the necessary condition of their functioning. Therefore, understanding of the intimate molecular mechanisms of MPs' action may be achieved only via exploring of helix-helix interactions in membrane-mimic media. This, in turn, requires that the spatial structure of TM protein oligomers is determined.

Delineation of MPs' structure using experimental techniques is significantly hampered by a number of difficulties concerned with the sample preparation, preservation of the native conformation, and so forth. This is clearly illustrated by the fact that MPs represent less then 1 % of 3D structures accumulated so far in the Protein Data Bank. Moreover, the only structure of TM helical dimer is available. Important new insights into the structure-function mechanisms of MPs and their complexes can be gained by the use of molecular modeling techniques. Fast estimation of TM helix association can be made via simulations in implicit membranes. Unfortunately, high computational efficiency is often attained at the cost of simplifications inherent in the membrane model. Such calculations therefore provide only rough model of membrane dimers. The refinement of these models can be done using MD simulations in all-atom hydrated lipid membranes and micelles. This approach is one of the most powerful computational techniques because it provides not only the information

about atomic-scale details of protein-membrane interactions but also takes into account the effects of proteins on structural and dynamic characteristics of membranes.

In this work we present the results of modeling of the spatial structure of a dimer formed by TM fragments of the protein BNIP3 using MD simulations in explicit DMPC bilayer. The pro-apoptotic protein BNIP3 belongs to the Bcl-2 superfamily, which is implicated in mitochondria-mediated apoptosis. A special attention is given to possibilities of MD approach in prediction and analysis of the dimeric structure. The quality of prediction is estimated by direct comparison of MD results with experimental structure, recently determined in our laboratory by NMR spectroscopy in DMPC/DHPC bicelles (Bocharov *et al*., manuscript in preparation).

METHODS

Starting conformations of BNIP3 dimers [sequence K_{163}VFLP SLLLS HLLAI GLGIY IGRR$_{186}$] were obtained using Monte Carlo (MC) conformational search in implicit membrane. Description of the implicit model and the related computational protocols were described earlier (Vereshaga *et al.*, 2005). In total, four different structural models of the dimer were obtained. All these conformations were relaxed in the presence of the implicit DMPC bilayers.

The spatial structure of the dimer obtained via MC-search was placed in the center of equilibrated DMPC bilayer (512 molecules). Then, lipid molecules intersected with the peptides, were removed. The resulting system was placed into rectangular box and solvated by SPC water molecules. The systems were then equilibrated by energy relaxation *via* 5×10^4 steps of steepest descent minimization followed by heating from 5 K to the temperature of simulations (315 K) during 500-ps MD run with the fixed positions of the peptides. After that, water molecules located in the bilayer interior were removed and the system was heated to the temperature of simulation during another 500-ps MD run without any restraints. Finally, the long-term (10 ns) collection MD run was carried out. Description of the simulation details is given in (Volynsky *et al*., 2005).

Analysis of MD trajectories was performed using original software developed by the authors and utilities supplied with the GROMACS package. The bilayer structure was characterized using the area per lipid molecule, the mean order parameter of acyl chains of lipids and the distance between phosphorus atoms of different monolayers. Stability of the monomer and the dimer was analyzed by calculation of the secondary structure of the peptides, root-mean-square deviation (RMSD) of their heavy atoms from the starting conformation, fluctuation of Cα atoms along the membrane normal. The packing of helices in the dimer structure was characterized by the distance d and angle θ between the helix axes, as well as by the contact surface area on the helix-helix interface. Interactions of the dimer with the environment were estimated in terms of interaction energy of its amino acid residues with lipid molecules and water. Hydrogen bonding of the peptide was also thoroughly explored. All these parameters were averaged over the equilibrium parts of corresponding MD trajectories (last 5 ns).

RESULTS AND DISCUSSION

Insertion of the BNIP3 dimers into the pre-equilibrated DMPC bilayer does not lead to noticeable distortions of the integrity of the lipid-water system. Adaptation of the dimer's structure to the membrane environment takes place mainly at the first stage of MD. It is accompanied by redistribution of protein interaction with water and lipid phases of the system. Geometry of the dimer, its interaction with the environment, and hydrogen

bonding equilibrate at the last stage of MD. All this data point out to correct choice of the configuration for the starting system and to adequate simulation protocol.

The first question that can be solved by MD simulation is the determination of the dimer structure. Analysis of the MD results shows that there are three possible packing geometries. Two of them correspond to a right-handed helical supercoil with the distance between the helical axes (d) ~ 7–8 Å, and the angle Θ ~ -30° (Fig. 1). The third model is a left-handed structure with d ~ 9 Å and Θ ~ 40°. Analysis of the time evolution (in MD) of geometric and energy characteristics of the dimer shows that the left-handed model is less stable – the distance between monomers grows up, being accompanied with the loss of intermonomer contacts. In our opinion, such analysis may be useful for discrimination between the native-like and misfolded conformations of the dimer.

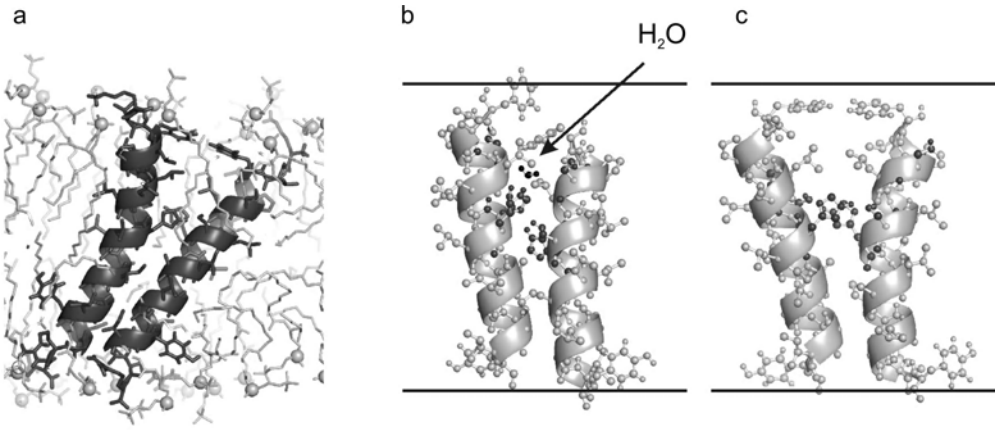


Figure 1. Results of MD simulation of the BNIP3 dimerization. *a* – equilibrium structure of the dimer in membrane. Protein is shown by the dark ribbon. Lipid molecules are displayed with gray sticks. Phosphorus atoms are presented by spheres. *b*, *c* – conformational variability of the dimer structure. Different models of dimer structure obtained in MD simulations. Polar residues are selected by dark gray. Lipids and water molecules removed for clarity. Positions of the membrane-water interface (the mean position of phosphorus atoms) are shown by black lines.

The dynamic characteristics of the membrane-dimer system are very important for determination of the structure-activity relationship for a given protein. Thus, analysis of fluctuations of BNIP3 residues along the membrane normal shows that the N-terminal part of the peptide is much more flexible. This corresponds to different H-bonding patterns between the peptide and the bilayer on its N- and C-termini. On the C-terminus Arg residues form up to ten long-living H-bonds with the lipids headgroups. In contrast, the N-terminal residues form only five short-living H-bonds. Another reason for such a flexibility consists in interaction between the monomers. In the region His173-Arg185 the helix-helix interface is very similar for the two right-handed models. It is formed primarily by small residues Gly and Ala. On the N-terminus the pattern of inter-monomer contacts is somewhat different depending on the model. In this case the interface consists of polar side chains of residues Ser168, Ser172, and His173. In this part of the dimer the helix packing is not as tight as on the C-terminus. As a result, the presence of the free volume and the existence of a number of donors and acceptors of H-bonds determine the structural variability of helical packing in the N-terminal part of the peptides.

The next question which can be addressed via MD simulations is the influence of the dimer on the bilayer properties. As mentioned above, insertion of the pair of helices do not disturb the whole bilayer structure. But some local changes are observed for the neighboring lipids. In particular, the mobility of these DMPC molecules is significantly restricted comparing with the rest of the bilayer. Another effect consists in penetration of

water molecules into the hydrophobic part of the membrane. Analysis of hydrogen bonding of the peptides with water shows that water molecules can reach the residue His173 (Fig. 1*b*), which is situated approximately in the middle of the membrane. It should be noted that such diffusion of water was detected only in the N-terminal part of the dimer. Therefore, this can not be an artifact of the simulations. Instead, such a phenomen is determined by the dimer structure.

And in conclusion, some notes should be made concerning the validity of the simulation results. To check their correctness, MD-data were compared with the experimental ones. Recently, the high-resolution structure of the TM dimer of BNIP3 in DMPC/DHPC bicelles was solved by NMR spectroscopy (Bocharov E.V. *et al*., manuscript in preparation). Inspection of the both (experimental and theoretical) models reveals that one of the predicted right-handed models is in a good agreement with the experimental one – they have similar packing and low (~1.5 Å) RMSD values. Moreover, NMR data point out to accessibility of the His173 residue to water – exactly the same conclusion was reached based on MD simulations. Also, NMR-derived models demonstrate some conformational exchange in the N-terminal regions of TM helices. Therefore, a proposal was made that the second predicted right-handed model can also be realized in the membrane-like environment.

ACKNOWLEDGEMENTS

This work was supported by the Program RAS MCB, by the Russian Foundation for Basic Research (grants Nos 04-04-48875-a, 05-04-49346), by the Russian Federation Federal Agency for Science and Innovations (The State contract 02.467.11.3003 of 20.04.2005, grants SS-4728.2006.4, MK-5657.2006.4). We thank Dr. Bocharov E.V. and Pustovalova Ju.E. for providing the NMR-structures of BNIP3 before their submission to the Protein Data Bank.

REFERENCES

Vereshaga Y.A., Volynsky P.E., Nolde D.E., Arseniev A.S., Efremov R.G. (2005) Helix Interactions in Membranes: Lessons from Unrestrained Monte Carlo Simulations. *J. Chem. Theory Comput*., **1**, 1252–1264.

Volynsky P.E., Polyansky A.A., Simakov N.A., Arseniev A.S., Efremov R.G. (2005) Effect of lipid composition on the "membrane response" induced by a fusion peptide. *Biochemistry,* **44**, 14626–14637.

Indexes

AUTHOR INDEX

KEY WORDS

Научное издание

Труды пятой международной конференции
“Биоинформатика регуляции и структуры генома”
Т. 1
на английском языке

Proceedings of the Fifth International Conference
on Bioinformatics of Genome Regulation and Structure
V. 1

Отредактировано и подготовлено к печати
в редакционно-издательском отделе ИЦиГ СО РАН
Редакторы: А.А. Ончукова, И.Ю. Ануфриева
Дизайн А.В. Харкевич
Компьютерная графика: А.В. Харкевич, К.В. Гунбин., Т.Б. Коняхина
Компьютерная верстка: А.В. Харкевич, К.В. Гунбин., Н.С. Глазкова

Подписано к печати 20.06.2006 г.
Формат бумаги 70×108 1/16. Печ.л. 29,8. Уч.-изд.л. 35,4
Тираж 250. Заказ 262

Институт цитологии и генетики СО РАН
630090, Новосибирск, пр. акад. М.А. Лаврентьева, 10
Отпечатано в типографии Издательства СО РАН
630090, Новосибирск, Морской пр., 2