RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

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

Volume 1

BGRS'2000
Novosibirsk, Russia
August 7-11, 2000

ICG, Novosibirsk, 2000

International Program Committee

Nikolay Kolchanov, Institute of Cytology and Genetics, Novosibirsk, Russia (Chairman of the Conference)
G. Christian Overton, Center for Bioinformatics, University of Pennsylvania, USA (Co-Chairman of the Conference)
Ralf Hofestadt, University Magdeburg, Germany (Co-Chairman of the Conference)
Patrizio Arrigo, Institute of Electronic Circuits, CNR, Italy
Martin Bishop, Human Genome Mapping Project Resource Centre, UK
Philip Bourne, SDSC, San-Diego, USA
Philipp Bucher, Swiss Institute for Experimental Cancer Research, Switzerland
Chris Burge, MIT Center for Cancer Research, Cambridge, MA, USA
Julio Collado-Vides, National University of Mexico, Mexico
Jim Fickett, SmithKline Beecham Pharmaceuticals, USA
Mikhail Gelfand, Institute of Protein Research, RAS, Moscow, Russia
Charlie Hodgman, GlaxoWelcome Research Medicine Center, UK
Minoru Kanehisa, Kyoto University, Kyoto, Japan
Kotoko Nakata, National Institute of Health Sciences, Tokyo, Japan
Leonid Kalinichenko, Institute of Problems of Informatics RAN, Moscow, Russia
Lev Kisselev, Engelhardt Institute of Molecular Biology, Moscow, Russia
Luhua Lai, Institute of Physical Chemistry, Peking University, Beijing, China
Hwa A. Lim, D'Trends, Inc, USA
Gerhard Michal, Tutzingen, Germany
George Michaels, Genomics Lead Development, Monsanto Co.USA
Luciano Milanesi, ITBA, Milan, Italy
Andrey Mironov, State Center for Applied Genetics, Moscow
Ken Nishikawa, Center for Information Biology, National Institute of Genetics, Japan
Manuel Peitsch, Glaxo Wellcome Experimental Research SA,Geneva, Switzerland
Mikhail Ponomarenko, Institute of Cytology and Genetics, Novosibirsk
Vadim Ratner, Institute of Cytology and Genetics, Novosibirsk, Russia
John Reinitz, Mt. Sinai Med. School, USA
Aida Romashchenko, Institute of Cytology and Genetics, Novosibirsk, Russia
Akinori Sarai, RIKEN Tsukuba Life Science Center, Tsukuba, Japan
Victor Solovyev, The Sanger Centre, Cambridge, UK
Masaru Tomita, Bioinformatics Laboratory of Keio University, Japan
Eduard Trifonov, Weizmann Institute of Science, Rehovor, Israel
Vladimir Tumanian, Engelhardt Institute of Molecular Biology, Moscow, Russia
Edgar Wingender, GBF, Braunschweig, Germany
Michael Zhang, Cold Spring Harbor Laboratory, Cold Spring Harbor, USA

Local Organizing Committee

Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk
Dmitry Afonnikov, Institute of Cytology and Genetics, Novosibirsk
Vasily Areschenko, Siberian Branch of RAS, Novosibirsk
Elena Borovskikh, Institute of Cytology and Genetics, Novosibirsk
Dmitry Grigorovich, Institute of Cytology and Genetics, Novosibirsk
Nadya Omelianchuk, Institute of Cytology and Genetics, Novosibirsk
Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk
Anatoly Kushnir, Institute of Cytology and Genetics, Novosibirsk
Anatoly Kurbatov, Institute of Archeology and Ethnography, Novosibirsk
Sergey Lavryushev, Institute of Cytology and Genetics, Novosibirsk
Yuri Orlov, Institute of Cytology and Genetics, Novosibirsk
Galina Orlova, Institute of Cytology and Genetics, Novosibirsk

INTRODUCTION

Two volumes of Proceedings of the International Conference BGRS-2000 encounting about 180 abstracts are aimed to direct an attention to the actual problems in bioinformatics of genome regulation and structure. The Conference BGRS-2000 organized by the Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences will be held in Novosibirsk, Russia, in August 7-11, 2000. This Conference will be the second in the series: the First International Conference on Bioinformatics of Genome Regulation and Structure – BGRS-98 was held in Novosibirsk in August 1998.

The question may arise: Why the Conferences BGRS attract their attention directly to the problems dealing with genome regulation and structure? The answer could be as follows: the structure and regulation of genome are the counterparts of life at molecular level; that is why understanding of fundamental principles of regulatory genomic machinery is impossible unless their structural organization is known, and *vice versa*.

During two years that have passed from the first BGRS Conference, the experimental genome study including applications to direct sequencing and mapping became of ever-growing scale. The huge bulk of experimental data on nucleotide sequences of complete bacterial genomes, the sequencing of which became a routine procedure in molecular biology, are being accumulated. Besides, complete genome of Drosophila is being deciphered, and human genome sequencing is drawing towards completion.

The ever-growing impact in genome studying is produced by novel experimental techniques. In particular, the EST technique is widely used in studying gene structure and gene expression patterns. Besides, microarray methods aimed at extracting unique and complete information on genome functioning and enabling to study simultaneously the expression patterns of dozen thousands of genes including those obtained at a single cell level, become implemented massively. In addition, single nucleotide polymorphism (SNP) technique provides a huge bulk of experimental data for studying regularities in mutation-assisted genome variability. Large-scale proteomic initiatives in the near future will lead to accumulation of large massifs of information on structure-functional organization of proteins.

The huge volume of experimental data that has been acquired on genome structure, functioning and gene expression regulation demonstrate the blistering growth. Development of informational-computational technologies of novel generation is a challenging problem of bioinformatics. Bioinformatics has entered that very phase of development, when decisions of the challenging problems determine the realization of large-scale experimental research projects directed to studying genome structure, function, and evolution.

By analyzing the papers submitted for publication in the two-volume issues of the BGRS-2000, the Organizing Committee came to a conclusion that participants of the Conference have concentrated their attention at consideration of the hottest items in bioinformatics listed below:

(1) Development of the novel generation of databases providing more complex, deep, and comprehensive description of (i) genome structure, function, and evolution, (ii) regulatory genome sequences, (iii) regulatory proteins, (iv) genetic networks, (v) signal transduction pathways and genetically controlled metabolic pathways.

(2) Development of computer technologies for automated knowledge discovery and data mining in the databases: ultra-rapid experimental methods developed for extracting molecular-biological data should correspond to similarly advanced technologies designed for automated treatment of these data, these technologies enabling to get the maximum of reliable and significant knowledge about genome function, regulation, and structure out of computer databases.

(3) Development of rigorous scientific methods for analysis of gene structure, discovery and modeling. Along with traditional approaches based on recognition of potential regulatory sites and coding regions, and their combinations, the more resolving power in solving this problem is demonstrated by the approaches based on comparative genomics, including both data search throughout databases and comparison of extended genome regions and even complete genomes (in case of bacteria).

(4) Development and improvement of methods in comparative genomics that became one of the most high-powered and perspective directions in modern bioinformatics. The efficient algorithms developed within the frames of comparative genomics appear to be more reliable tool acquired for gene recognition and gene reconstruction throughout *de novo* sequenced genome DNA, for recognition of regulatory elements controlling genome functions and gene expression regulation. Besides, comparative genomics will go a long way towards revealing fundamental principles of genome organization and regularities in genome molecular evolution.

(5) Further mastering of approaches designed within the frames of comparative genomics strongly depends upon comprehension of fundamental regularities in genome organization and evolution. That is why computer analysis and modeling of genome mutability, together with studying of fundamental laws of evolution of genomes, coding gene regions and regulatory genomic sequences become the matter of especial importance. Accumulation of knowledge in this field will certainly help in searching for objective methods in annotating and finding of genes and regulatory signals in genomic sequences.

(6) Development of novel generation of mathematical algorithms implemented for analysis of regulatory genome sequences (RGS) and for accounting of real complexity of RGS. These algorithms are characterized by a large variety of parameters significant for gene functioning, by blockwise structure, and hierarchy in RGS organization. On the background of these algorithms, the fine accuracy methods are being developed for

recognition and prediction of quantitative values of regulatory genomic sequences activity of various types, which provide implementation of numerous genome functions regulating basic stages of gene expression.

(7) Revealing of fundamental regularities in structure-functional organization of RGS controlling basic types of molecular-genetical processes (i.e., replication, transcription, splicing, polyadenylation/processing, translation, etc). Besides revealing the regularities in structure-functional organization of RGS that are valuable for increasing the accuracy of their recognition, this analysis allows to obtain a fundamental knowledge on molecular mechanisms of RGS functioning, thus enabling to solve one of the main problems in bioinformatics of genome regulation and structure.

(8) Development of methods aimed at prediction and recognition of structure-functional organization of proteins encoded by the genes detected within *de novo* sequenced genome sequences. The lack of unified technological production line processing from the coding gene regions in the sequenced genomes to prediction of structure-functional organization of proteins encoded by these genes serves as the stopping brakes for implementation of large-scale genome projects. It should be stressed that during the solving of the task, a large attention should be paid to detecting fundamental principles of protein organization and evolution. The most important is the studying of aspects of protein function and structure related to genome regulation. During the recent years, the tendency manifested itself in convergence and intersection of the lines in bioinformatics of genome regulation and structure and in computer-assisted proteomics. This observation is clearly approved in Proceedings of BGRS-2000.

(9) Large-scale genome analysis. The other day computer analysis was restricted to studying of local context regularities in genome structure. Currently, due to widespread sequencing of complete genomes and their extra-extended fragments, a possibility first appeared to analyze large-scale context dependencies in genome DNA organization. To tackle this problem, it is necessary to develop operative methods aimed at analysis of extra-extended genome sequences.

(10) Description in databases and modeling of genetical networks, which control the processes of basic metabolism, cell division and differentiation, organ- and tissue morphogenesis, growth and development of an organism; support of homeostasis of molecular, biochemical, and physiological parameters of organisms, etc. Systemic investigation of mechanisms related to genome functioning and gene expression regulation at the level of gene networks and signal transduction pathways should be provided. On the grounds of these very processes, the key problem in bioinformatics, that is, recognition of phenotypical characteristics of an organism on the basis of information encoded in their genomes will be solved in future.

(11) Development of efficient technologies for integration of informational and software resources on the structure and regulation of genomes and designing on this basis of super-large computer systems implemented for analysis and modeling of intricate molecular-genetic systems and processes.

(12) Analysis of fundamental regularities in (i) genome functioning, organization, and evolution, (ii) the mechanisms governing the coding of genetical information, (iii) molecular bases of realization of genetical language, principles of organization, functioning, and evolution of genetical networks and molecular-genetic systems.

All the questions listed above will be suggested to consideration of participants of BGRS'2000 at 7 sections and presented in a form of plenary lectures, oral communications, posters, Internet computer demonstrations and round table discussions.

BGRS'2000 will bring together the experts in Bioinformatics to discuss the progress in the field of bioinformatics of genome regulation and structure achieved at the end of 20th century, the basic approaches devoted (i) to data description and analysis; (ii) modeling of complex molecular-genetical systems; (iii) to revealing of fundamental principles of genome organization and evolution and of mechanisms of genetical information coding; (iv) to evaluation and marking off the future trends in this field.

The researchers working in the fields of experimental biology and interested in application of Bioinformatics methods in their work are also the participants of the Conference. With this respect, the Conference is expected to be a stimulating event not only giving a future development of bioinformatics as it is, but also establishing new links between Bioinformatics and experimental research.

By working out the BGRS2000 schedule, the Organizing Committee has tried to keep the balance between technical (applied) and fundamental aspects in bioinformatics. This principle has a clear reflection in contents of Proceedings of the Conference. Herein, we have tried to follow the well-known and far-back principle: «nothing is more practical than the good theory».

Professor Nikolay Kolchanov
Co-Chairman of the Conference
Head of Laboratory of Theoretical Genetics
Vice-Director of the Institute of Cytology and Genetics
Novosibirsk Russia

Professor Ralf Hofestadt
Co-Chairman of the Conference
University Magdeburg, Germany

G. Christian Overton

15.02.1948 – 1.06.2000

The death of Dr. Overton, premature and unexpected, is a body blow for all Bioinformatics community and for us, his friends and colleagues in Russia. Chris always paid a great attention to strengthening of the international cooperation in the Bioinformatics research. That is why, he left behind along with brilliant scientific results an example of real international collaboration in science. As an example of international cooperation initiated by Chris may serve the collaboration with Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of Russian Academy of Sciences. From 1995, this laboratory and the Center for Bioinformatics at University of Pennsylvania together made a common research despite of distance between Philadelphia and Novosibirsk, and we are very grateful to Chris for his support, assistance and understanding friendly provided.

Dr. G. Christian Overton was the founding Director of the Center for Bioinformatics at Penn, established in 1997 as an interdisciplinary venture between the Schools of Medicine, Arts and Sciences, and Engineering and Applied Science. He was also an Associate Professor in the Department of Genetics, and held a secondary appointment in the Department of Computer and Information Science in the School of Engineering and Applied Science. Dr. Overton received his Bachelor of Science degree in Mathematics and Physics from the University of New Mexico in 1971, his Ph.D. in Biophysics from the Johns Hopkins University in 1978, and his M.S.E. in Computer and Information Science from the University of Pennsylvania in 1986. After receiving his M.S.E, he returned to the University of Pennsylvania in 1991 as an Associate Professor. In addition to his research, Dr. Overton was an Editor for the Journal of Computational Biology, Bioinformatics, and Gene/Gene-COMBIS as well as the Member of the Board of Directors for the International Society for Computational Biology.

Dr. Overton's brilliant skills in biophysics and bioinformatics, his deep understanding of the challenges in biology, medicine, and computer science enabled him to organize many outstanding research projects, which bridge the gap between experimental biology and computer science aimed to experimental data treatment. Dr. Overton is internationally recognized as a pioneer in genomic research and application of computational approaches for solving biological problems. He focused on problems associated with database integration, genome annotation, gene recognition, and detection of regulatory elements governing the expression of many genes that comprise the human genome.

Chris was one of the Co-Organizers of the BGRS2000 Conference. Due to his activity, the Organizing Committee managed to put together the effort of those interested in the basic approaches and trends in bioinformatics. Chris will be remembered for his love of science, his charm, good nature and his great intelligence. For people here in Novosibirsk who met him and knew him well he will be remembered as a faithful and good friend and as a researcher devoted to science.

Professor Nikolay Kolchanov,
Head of Laboratory of Theoretical Genetics,
Vice-Director of the Institute of Cytology and Genetics,
Novosibirsk, Russia

CONTENTS

BGRS 2000

SECTION 1. BIOINFORMATICS OF REGULATORY GENOMIC SEQUENCES

INTEGRATED SYSTEM ON GENE EXPRESSION REGULATION GENEEXPRESS - 2000

Kolchanov N.A., Podkolodny N.L., Ponomarenko M.P., Ananko E.A., Ignatieva E.V., Kolpakov F.A., Levitsky V.G., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Vorobiev D.G., Lavryushev S.V., Grigorovich D.A., Ponomarenko J.V., Kochetov A.V., Orlova G.V., Kondrakhin Y.V., Titov I.I., Vishnevsky O.V.,Orlov Yu.L, Valuev V.P., Ivanisenko V.A, Oschepkov D.Yu., Omel'yanchuk N.A., Pozdnyakov M.A., Kosarev P.S., Goryachkovskaya T. N., Fokin O.N., [+]Kalinichenko L.A., [+]Kotlyarov Yu.V.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
+Institute for Problems of Informatics RAS, Moscow, Russia
e-mail: kol@bionet.nsc.ru
*Corresponding author

Keywords: gene, genome, genenetworks, regulation, expression, transcription, splicing, translation, DNA, RNA, proteins, regulatory sequences, databases, integration, knowledge discovery

Availability:

http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/

1. Problems of molecular-biological information resource integration

A discriminative feature of molecular-genetic systems is their complex hierarchical and/or network organization. For instance, an organ is consists of tissues, a tissue – of various cell types, a cell – out of compartments (i.e., cytoplasm, nucleus, vacuoles, etc.) that contain the macromolecules of DNA, RNA, and proteins. These macromolecules intensively interact with each other (they organize complexes, act in various reactions, move through cell compartments, cells, tissues, and organs, etc.), thus forming a composite net of interactions, namely, the gene network.

While solving concrete problems that are important in practice it is necessary to use a large number of heterogeneous, weakly structured molecular-genetical databases accumulating the results of numerous, complementary, intersecting and probably contradictory experimental data. Databases on molecular-genetic information store the sequences, structures, 3D descriptions, attributive information, along with program software tools for data analysis, search of regularities, and prediction of different properties of objects, data reorganization, visualization, etc.

Among the most actual are the tasks that are characterized either by scarce data, or by the data, which are difficult to compare due to their heterogeneity. As heterogeneity, we understand here not only differences in formats and the ways of data representation, but also semantical heterogeneity of information, which should be integrated into a system.

In particular, one of the problems is the difference of objects that a user could consider as homogeneous under solving a particular problem. Homogeneous objects, in turn, could have different attributes, which show only indirect evidence with some accuracy about the knowledges on an object that are needed for a user. Even one and the same characters could be measured by different methods, with different accuracy, in various conditions. All these facts need to use adequate approaches based on semantic analysis for comparison and integration of data obtained from various sources [1].

For integration of resources on gene expression regulation, the super-large system GeneExpress is being developed at the Institute of Cytology and Genetics of SB RAS. This system integrates a large bulk of databases, and hundreds of programs for treatment of information on structure and function of DNA, RNA, and proteins [2].

2. Methods for molecular-biological information resource integration, used in the GeneExpress system

For integration of heterogeneous molecular-biological information resources (IR), we are developing the following approaches (Fig. 1):

- Hypertextual integration of IR. For this purpose, we use the standard Web-technologies, tools for automated generation of hyperlinks providing relationships between different IR and linking between databases via key fields;
- Development of unified object-oriented environment on the base of data mapping into canonical models by means of wrappers and mediators;

- Semantic data integration based on ontology. It includes automated verification, formal and justified from the view of biology and directed to test the data integrity, correctness, and consistency;
- Data mining, based on their automated processing and goal-seeking transformation, search of dependencies, extraction of the samples, and construction of daughter databases;
- Development of common query system, which will provide simultaneous queries addressed to different databases, and usage of Thesauruses and ontology libraries for the automated query management to the databases. The system for generation of scenarios for data analysis and processing.

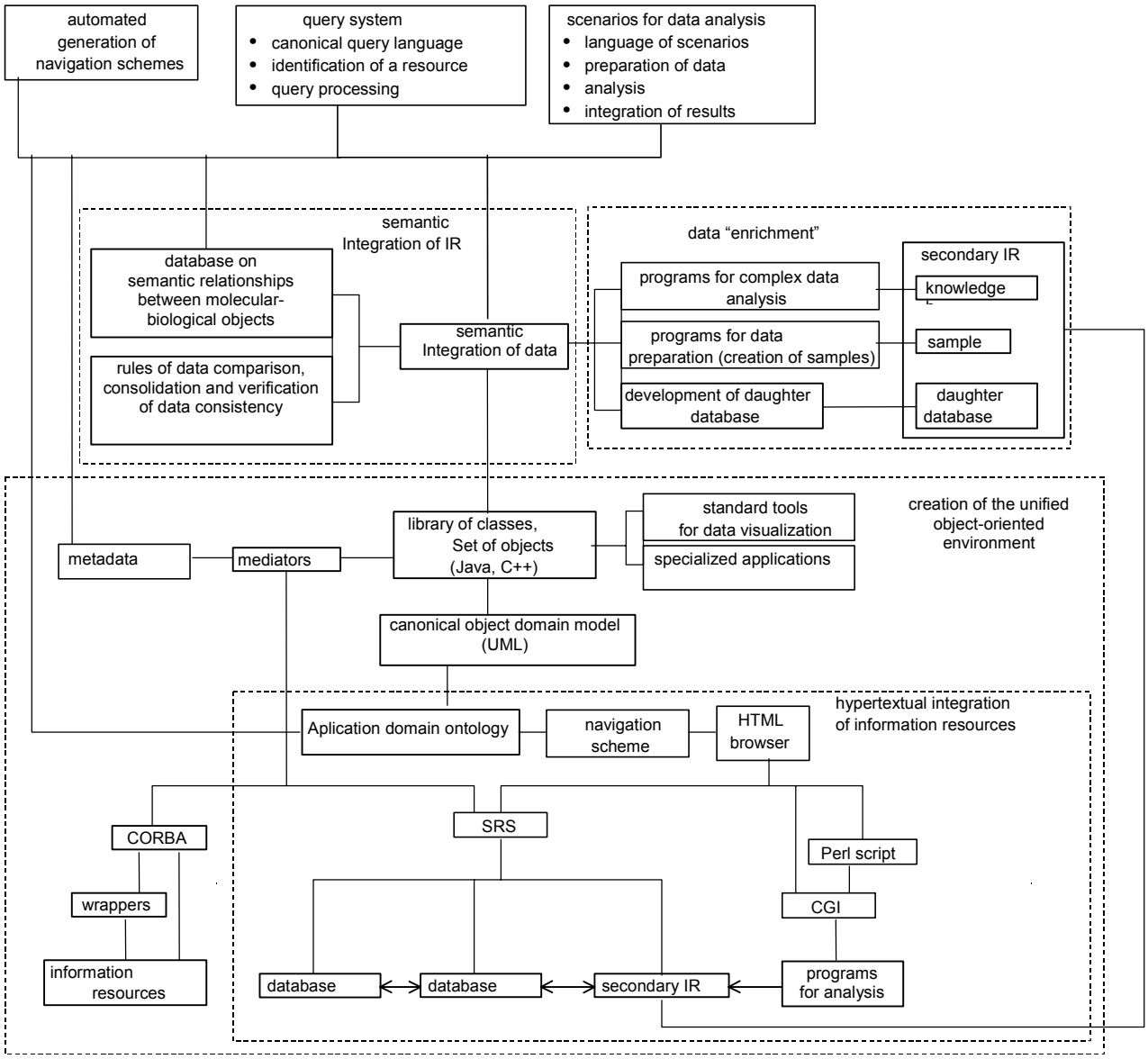


Figure 1. Integration of molecular-biological information resources in the system GeneExpress.

Following the approach suggested, one may discriminate 3 levels of the system: (i) a user's level, federative level, and the level producing an access to information resources. At the level of access to non-homogeneous IR, the special agents, mediators are used. They produce transformation of information from the databases into the set of objects corresponding to canonical model, that is into a unified for all the data representation. The federative level includes the tools for semantic analysis and data integration, the tools for data processing, recognition, generation of the daughter databases and generation of knowledge on the basis of automated search of regularities, etc. The user's level includes the agents, or mediators, which provide data reorganization in accordance to the user's demands.

Among the traditional approaches applied by the authors is the hypertextual integration between the databases and the programs for data analysis developed in the IC&G (Table 1), along with the other publicly available

Internet resources on gene expression regulation. An access to these databases is provided by the SRS (Sequence Retrieval System).

For the network access, the SRS system uses the standard CGI interface for the WWW server. Query manager enables to combine the queries as a logic conditions of an arbitrary complexity. The query result is produced in a form of the HTML-file. Besides, SRS is supplied by flexible tools for rearrangement of data in the course of their representation as an HTML-document, by the tools for automated generation and inclusion into hypertextual documents of clear hyper-links to the other documents or queries to databases, along with incorporated standard formats for representation of molecular-genetical data. By accounting these positive features, the SRS is used as one of the main tools for data integration in the GeneExpress system.

For an access to the programs of molecular-biological data analysis, we use a CGI interface and Perl scripts, which are destined for the treatment of the HTML-forms, data input and output, programs for reformatting of the data introduced for analysis into the format required by the programs for analysis, etc.

The functional scheme of the GeneExpress is given in Fig 2. Each module contains: 1) experimental data represented as a database or some sample; 2) program for data analysis; 3) results of an automated data processing; 4) tools for the graphical representation of these data and the results of the data analyses.

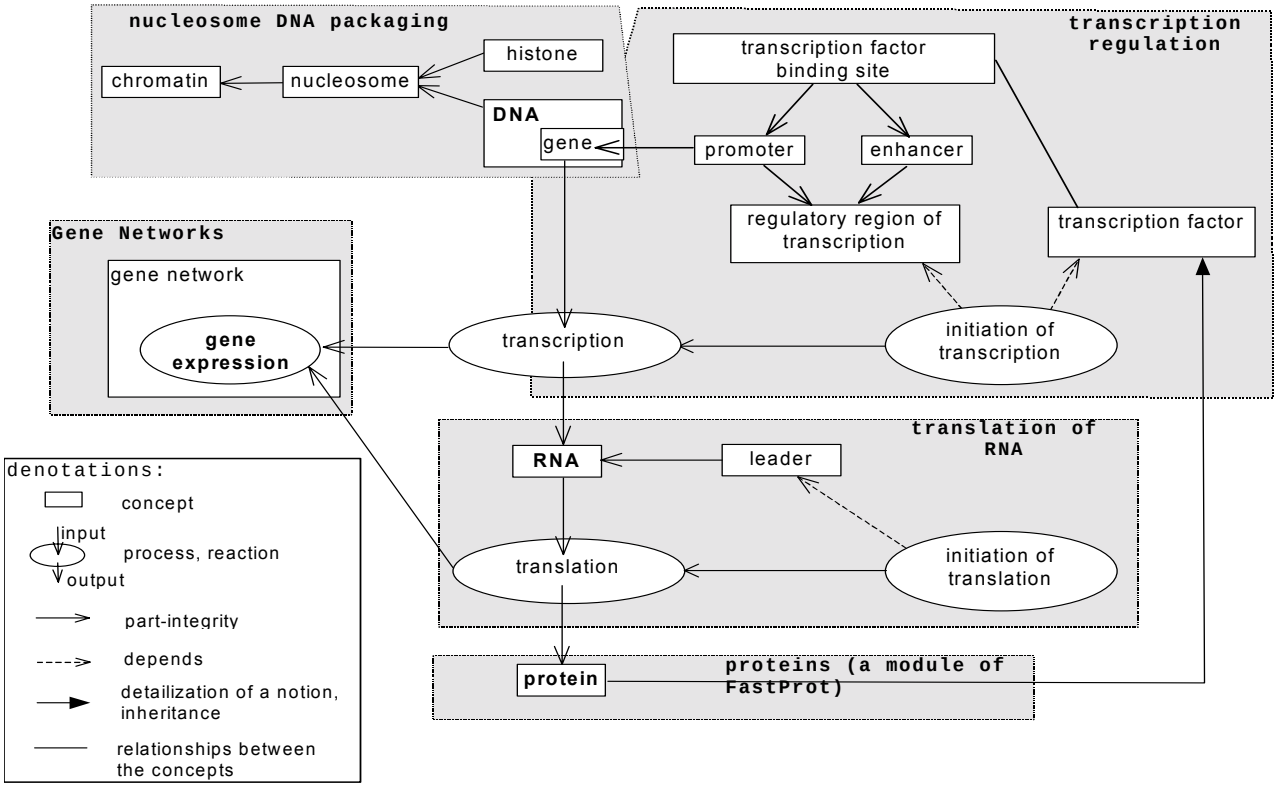


Figure 2. Correspondence of the GeneExpress system modules (grey rectangles) to ontology of gene expression.

The subdivision into functional modules is based on the ontology of the object environment (gene expression regulation). This ontology describes conceptualization of a definite field of knowledge. It includes such notions as a gene, regulatory region, transcription factor, site, organism, organ, tissue, stage of development, etc. We use the following types of relationships: general-especial, content, classification, origin, function, regulator of a function, relationships of location, participation in a process, temporal relationships, and so on.

For immersion of molecular-biological databases available via SRS into the object-oriented environment, we are developing the mediators, which provide the parsing of information from the databases. These mediators display the information as a set of objects of Java classes corresponding to the general (canonical) model, within the frames of which integration of heterogeneous data proceeds.

At the current moment, such mediators are created for the databases TRRD [3], GeneNet [4], and for the EMBL databank, which accumulates all decoded sequences of DNA and RNA molecules.

For the main types of molecular-biological data within the canonical model, we have developed the tools for their graphical representation, these tools are given as the library of classes in Java. This permits to visualize by the standard way an information extracted from different databases and the results of analysis. We have realized two types of graphical representation of molecular-biological data:

1) a diagram (graph) is the way to represent organization of molecular-biological systems (e.g., gene networks, metabolic pathways, signal transduction pathways, etc.);

2) a map is the way to represent the data on structure-functional organization of sequences of gene clusters, gene sequences, of transcription regulatory regions, RNA, and proteins.

As examples of applications applying the library, the Java-applets TRRD Viewer and GeneNet Viewer, the graphical interface for data input into GeneNet database through the Internet [4], may serve.

We develop the approaches for creation of autonomously adjusted mediators that use the base of meta-information, which will include various types of conceptualizations, including the basic and object-oriented meta-knowledge, Thesauruses, meta-description of informational resources and types of access to these resources, to the methods for solving the problems, and to applications. The most important stage of this work is a formalized description of elementary structures, events, and processes significant for gene expression regulation (from the level of molecular events to the level of an organism), description of a hierarchy of the subject environment, relationships between them and etc.

The meta-information of the federative level of a mediator is supposed to include the following components:

- *Ontology of gene expression regulation* defines such notions as a gene, regulatory region, factor, site, organism, organ, tissue, stage of development, etc. As the conceptual relationships we use the notions like general-especial, content, classification, origin, function, regulator of function, relationships of localization, temporal relationships, and so on.
- *Terminological information* includes rubricator and Thesaurus, determining the sets of lexical units applied in the field of gene expression regulation.
- *Structure meta-information* includes the specifications of data types and classes, which are characteristic for information representation in the subject environment considered. Besides, the interpretation is given, it describes the components of schemes within the frames of ontological definitions and Thesaurus of a mediator.
- *Meta-description of the methods* that are applied for the treatment of information and for solving the tasks. Here we determine both the signatures of methods bound to their ontological semantics and specifications of the algorithms.
- *Meta-description of tasks* that are considered as typical processes of the task solving within the application environment of a mediator. This meta-information is represented as the specifications of typical flows of processes characteristic for the subject environment.

This library of ontologies will be used by the tools for semantic data integration and by the system of queries, which provide simultaneous query management to different databases.

The construction of this library of ontologies will enable to extent flexibly the formats of gene expression description, will make effective possibilities for integration of molecular-biological databases, and provide standardization of the notions under usage. Besides, these notions will be put in correspondence to the norms used in a subject environment considered. In addition, it will provide possibilities, based on ontology, to search information in the databases, and will provide more reliable control for the quality of information accumulated in the databases, etc.

Currently, we have developed the controllable vocabularies that include the notions used in the databases. The work is going on developing of a Thesaurus on gene expression regulation and ontological description of many notions used in the field of gene expression regulation.

Development of the tools for supporting the database of the meta-information including ontological descriptions and Thesaurus is done in the Oracle 8 environment installed at super-computer RM600/E30. To get an access to the databases on molecular-genetical data on gene expression regulation developed in the Institute of Cytology and Genetics SB RAS, we develop special adapters, repositories of interfaces and object's realizations, services. They are developed for the work with the objects, which provide registration of objects, generation and interpretation of objects, activation and deactivation, call-in of methods, etc. The relation version of the TRRD database is realized in the Oracle 8 environment.

3. Information resources of the GeneExpress system

GeneExpress is designed for accumulation of experimental data, data navigation, data analysis, and analysis of dependencies in the field of gene expression regulation. It integrates the large amount of databases and hundreds of programs for processing the data on the structure and function of DNA, RNA, and proteins, together with the other IR important for gene expression description and available via the Internet.

Transcription Regulatory Regions Database (TRRD) [3] describes the modules of transcription regulatory regions and the hierarchy of their organization: cis-elements, composite elements, promoters, enhancers, silencers, and the extended transcription regulatory regions. It provides description of different features of transcription regulation, namely, its dependence on the cell cycle stage, developmental stage, tissue-specificity, or effects of external factors, etc. It is linked to the relevant databases on transcription regulation (TRANSFAC, EPD, EPODB, COMPEL, GERD) as well as to various software for analysis of regulatory genomic sequences

included into the GeneExpress system. At present, GeneExpress contains the description of more than 950 genes including about 5800 transcription factor binding sites and different types of regulatory elements. The most detailed description is given for the cell cycle gene family, erythroid-specific genes, genes involved in lipid metabolism, interferon-inducible genes, glucocorticoid-regulated genes, muscle-specific genes, endocrine System genes, plant genes, heat shock-induced genes and some others.

Table 1. Main databases on gene expression regulation entering the GeneExpess system.

Databases	Brief description
TRRD	Transcription Regulation Regions Database. It contains an information about the structure and function of DNA regions governing gene expression regulation.
GeneNet	It contains a graphical representation of gene networks and description of elements, which compile these graphs: cells, genes, RNA, proteins, various chemical agents, together with description of interaction between all the components given above. 22 Gene networks.
Selex	It accumulates the sequences of various functional DNA regions detected by a definite experiment (Selex protocol). It accumulates 3311 selected randomized DNA/RNA sequences and 224 computer programs (C- codes) for site recognition.
Activity	A distributed and intelligent database for the activities of the functional sites in DNA and RNA. It describes 6819 site variants with 6836 activity values, 49 program codes (20 prediction, 29 - sequence-dependent conformational features), 245 references, etc.
Property	It accumulates an information on conformational and physicochemical properties of double DNA helix.
Leader	It stores the samples of particular RNA regions (leaders) that are responsible for initiation of translation. 869 Site Variants.
Samples	The database stores 63 samples including 7421 regulatory genome sequences of different types.

4. Data mining in the GeneExpress system

A discriminative feature of molecular-genetical experimental data is a necessity of their automated analysis, in particular, automated genome annotation, that is, isolation of functionally significant genome regions and prediction of their structure and function. To this aim, the approaches like data mining [5] are used. With this respect, GeneExpress system includes a large number of procedures for data treatment that are used for data mining that is, their step-by-step reorganization into more informative, from the user's point of view, semantic environment. In this process, each subsequent level uses the preceding semantic environment as an initial one. The direction of such multi-level "enrichment" of data depends upon the concrete task.

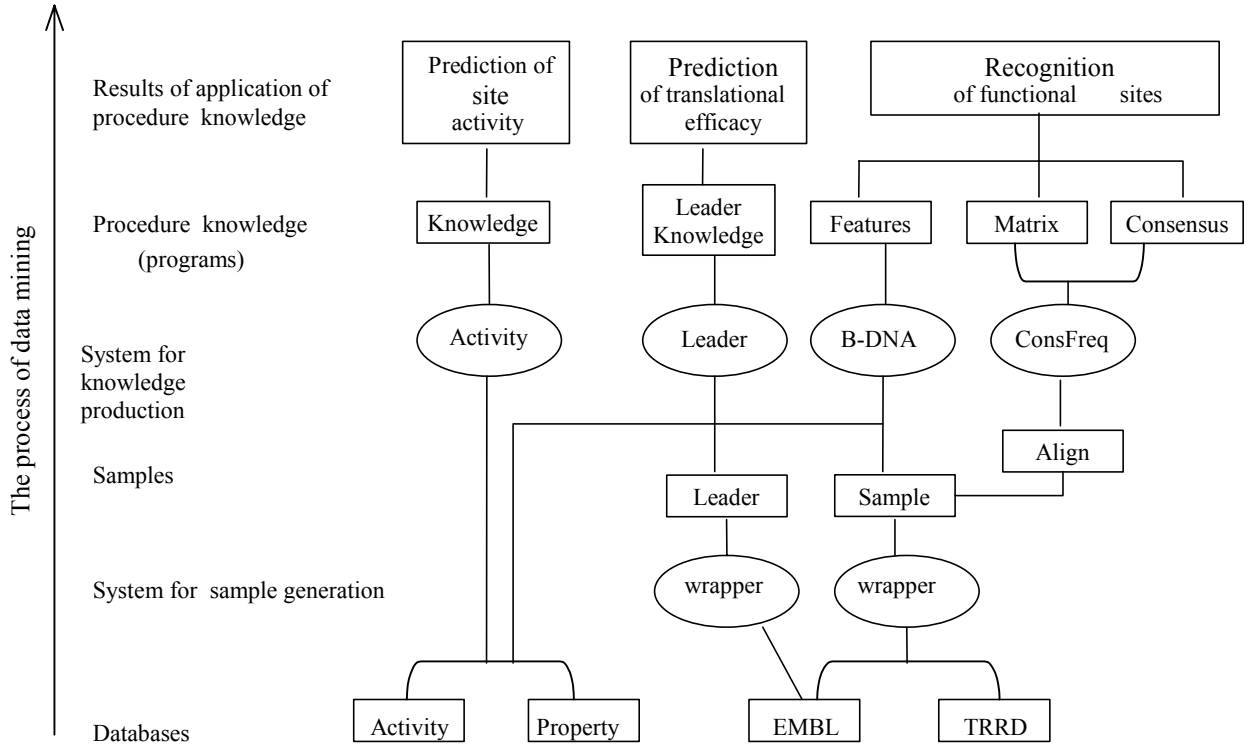


Figure 3. An example of application for data «enrichment» technology within the GeneExpress system.

At the first step of data mining (Fig. 3), an information (samples) is being prepared for automated production of knowledge. In order to create such samples, the special tools providing automated generation of samples from different regulatory regions (for example, leader sequences) are applied on the basis of semantic analysis of description of their structure-functional organization in the EMBL databank. Besides, the sequences of transcription factor binding sites may be generated automatically on the basis of integration of information from the databases TRRD and EMBL.

At the following stage, the systems for knowledge production (Table 2) produce automated treatment of the data extracted from the databases and/or from the samples constructed. Then the search for regularities significant for prediction of activity or the search of functional sites is made. The resulted procedure knowledge include (i) the procedure (program, script) necessary for a definite site type recognition of for activity prediction, (ii) description of the procedure application, (iii) conditions of its application, (iv) format of input data, (v) restrictions for input data, (vi) format of output data, etc. These knowledge enable to synthesize automatically complex scenarios for the search of different functionally significant regions within annotated genome sequences and to predict there structure and function.

Table 2. Systems for knowledge production and the resulted procedure knowledges, contained in the GeneExpess.

System for knowledge production	Resulted procedure knowledge (programs)
B-DNA provides production of knowledge on conformational and physico-chemical characteristics of sites*, significant for their functioning and recognition	Features – 1402 programs for site recognition by their significant conformational and physicochemical properties
Activity provides production of knowledge on context, conformational, and physicochemical properties of sites significant for prediction of their activity	Knowledge - 49 programs for site activity prediction in accordance with their nucleotide sequences
CONSFREQ is designed for production and usage of knowledge on context site properties, significant for these sites recognition	Matrix - 567 programs for site recognition by the frequencies of short "words"-oligonucleotides. Consensus includes 66 programs for site recognition by their evolutionary-conserved invariants
Leaderprovides production of knowledge on translation efficacy of various leader RNA sequences	Leader Knowledge accumulates the features of leader RNA sequences significant for their translation efficacy. It contains 96 program C-codes for prediction of the functional site activities in mRNA with the Motivations and references to experimental publications

*site – functionally important region of DNA or RNA macromolecule

Conclusion

The GeneExpress system developed by us includes a large number of databases and hundreds of programs for treatment of information on structure and function of DNA, RNA, and proteins. Application of contemporary approaches enables to provide an integration of all heterogeneous information and software resources on gene expression regulation and to create a qualitatively novel effective tool aimed at research in the field of gene expression regulation.

Acknowledgments

This work is supported by Integration Project of SB RAS IG2000/65, Russian Foundation for Basic Research (grants Nos 98-07-91078, 98-07-90126, 99-07-90203, 00-07-90337), National Human Genome Program, Integration grant of SB RAS No 66, and The Committee of Science and Technology of the Russian Federation.

References

1. Kalinichenko L.A. Methods and tools for non-homogeneous databases integration. M.: "Nauka", 1983 (in Russian).
2. Kolchanov N.A., Ponomarenko M. P. et al. (1999) Integrated databases and computer systems for studying eukaryotic gene expression // Bioinformatics. V. 15, № 7, P. 669-686.
3. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., et al. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. Nucleic Acids Res., 28, 298-301.
4. Ananko E.A., Kolpakov F.A., Kolchanov N.A. (2000) GeneNet database: a technology for a formalized description of gene networks// This issue.
5. Kolpakov F.A., Ananko E.A., Kolesov G.B., Kolchanov N.A. (1998) GeneNet: a gene network database and its automated visualization // Bioinformatics. V.14. P.529-537.
6. Vityaev E.E., Podkolodny N.L., Vishnevsky O.V. et al (2000) Detecting patterns of structure-function organization of regulatory genomic sequences in a first order logic// This issue.
7. Kochetov A.V., Ponomarenko M.P., Frolov A.S. et al (1999) Prediction of eukaryotic mRNA translational properties// Bioinformatics. V. 15, № 7, P. 704-712.
8. Ponomarenko M.P., Ponomarenko J.V., Frolov A.S. et al (1999) Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins// Bioinformatics. V. 15, № 7, P. 687-703

TRANSCRIPTION REGULATORY REGIONS DATABASE (TRRD)

Ananko E.A., Podkolodnaya O.A., Ignatieva E.V., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Stepanenko I.L., Goryachkovskaya T.N., Podkolodny N.L., Grigorovich D.A., Naumochkin A.N., Korostishevskaya I.M., Lokhova I.V., Romashchenko A.G., Kolchanov N.A.

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

Keywords: transcription regulation, database, database integration

Resume

Motivation:

Most of specialized databases accumulating information on gene transcription deal with particular aspects of this problem. In the frameworks of these databases, it is impossible to reveal the mechanisms of integral transcription regulation. A database TRRD was developed for accumulation of experimental data on the aggregate of molecular mechanisms of gene transcription regulation in eukaryotes together with peculiarities of the structural and functional organization of extended regulatory regions of these genes.

Results:

Current release of TRRD comprises the description of more than 900 genes and over 4200 transcription factor binding sites. This information was obtained through annotating more than 3000 scientific publications.

TRRD is installed under SRS that provides possibilities for making interactive queries to the database and its integration with other databases. The TRRD is represented as five tables containing information on a gene expression, binding sites and transcription factors.

Availability:

TRRD is available through the WWW at http://www.bionet.nsc.ru/trrd/.

Introduction

Successful research in the fields of molecular biology needs an extensive bulk of factual evidence accumulated in publications on gene expression regulation. Gene expression in eukaryotes is an extremely complicated biological process including several stages. The key events of expression regulation occur at the level of transcription.

In the context of an abundance of experimental data on gene expression and the huge rate of accumulation of the novel information, an access to these data is physically unfeasible without utilization of database technology. A great number of molecular biological databases have been developed containing information on structure and function of regulatory genomic sequences as well as programming tools for the data analysis, classification, and sorting. The most known are the databases accumulating general information on primary DNA structure, EMBL [Baker et al., 2000] and GenBank [Benson et al., 2000]; on primary protein structure, SWISS-PROT [Bairoch and Apweiler, 2000] and PIR [Barker et al., 2000]. Among specialized databases are: the database on transcription factors, TRANSFAC [Wingender et al., 2000], database on composite regulatory elements, COMPEL [Kel-Margoulis et al., 2000], database on eukaryotic promoters, EPD [Perier et al., 2000]. As a rule, listed databases are related only with discrete aspects of transcription regulation. In the frameworks of these databases, it is impossible to reveal the mechanisms of integral transcription regulation.

For accumulation of experimental data on structure and functional organization of extended regulatory regions in eukaryotic genes, a database TRRD was developed [Kolchanov et al., 2000].

Unlike the databases listed above, TRRD accumulates information about the whole integrity of molecular mechanisms of transcription regulation. Due to this feature, TRRD may serve as a basis for integration of informational resources devoted to studying molecular mechanisms of gene transcription.

Methods and algorithms

The TRRD database allows to accumulate the information on structural and functional features of gene regulatory regions, which may contain various tissue- and stage-specific regulatory elements, alternative promoters (and the corresponding alternative transcription start sites), silencers, enhancers, etc.

The program TRRD-INPUT is used to standardize the information in TRRD. The controlled vocabularies for database sections, supported by this program, have been considerably expanded. The vocabularies describing organs, tissues, and cells where the genes described in TRRD are expressed were essentially developed. Now these vocabularies are united and organized hierarchically. This hierarchy is used for modification of the queries (generalization or specification) and for realization of associated search in TRRD.

For installation of the TRRD database under the SRS (Sequence Retrieval System, developed and supported by the EMBL group) [Etzold et al., 1996], all the information contained in the flat file is divided between 5 interconnected tables (Table 1).

Table 1. TRRD SRS-tables*. An example of description of the human von Willebrand factor gene.

TRRD SRS-table	Information content	Entry example
TRRD-GENES	General information about the gene: identifier (**ID**), accession number (**AC**), annotator (**CR**), species (**OS**), gene name (**SN**, **NG**, **SY**), nucleotide bank (**BI**) and other databases (**DR**) links, key words (**KW**), chromosomal localization (**CH**), regulatory region description (**RG**, **PR**), site links (**PR**).	**ID** Hs:VWF **DT** 19/11/99 **AC** A00917 **CR** O.A.P. **OS** human, Homo sapiens **SN** VWF **NG** von Willebrand factor gene **SY** F8VWF **BI** EMBL; HSVWF123; X06828; ST: 836 **DR** SWISS-PROT; VWF_HUMAN; P04275; **KW** glycoprotein, adhesion protein, hemostasis **CH** 12 p13.3-p13.2 **RG** 5'region **PR** negative regulatory region; ST; -487 to -312; S4237 **PR** negative regulatory element; ST; -147 to -89; S4238 **PR** minimal promoter region; ST; -90 to +22; **PR** endothelial-cell specific positive regulatory region; ST; +156 to +234; S4236
TRRDEXP	Peculiarities of the gene expression: gene link (**ID**), organ (**RO**), tissue (**RU**), cells (**RL**), expression level (**RL**), environmental factor (**RI**) and its influence on the gene expression (**FF**), time of the induction (**RH**), and bibliography link (**TR**).	**RE** A00917.002 **ID** Hs:VWF **RT** mRNA **RO** vein **RU** endothelium **RN** endothelial cells **RL** present **RI** X-irradiation **FF** induction **RR** [Jahroudi N. et al., 1996a]
TRRD-SITES	Binding sites of transcription factors: site name (**NM**), gene link (**ID**), factor link (**TF**), influence to the gene transcription (**AT**) sequence (**SQ**), sequence positioning (**PQ**), footprint positions (**PF**), nucleotide bank (**BF**) and other databases (**DR**) links, codes of the experiments confirmed the site functionality (**AG**), and bibliography links (**AG**).	**AN** S4238 **ID** Gene: Hs:VWF **NM** Oct-1 bs; **DR** SAMPLES; OCT; **TF** Oct-1; **AT** decrease **SQ** ggcagTTAATTAAaggc **PQ** -137 to -121 **PF** -133 to-126 **BF** EMBL: X06828:698 **AG** Human umbilical vein endothelial cells, Dami, HeLa: 1.1.1 [Schwachtgen J.L. et al., 1998] **AG** Human umbilical vein endothelial cells, HeLa: 3.1, 3.3, 3.4, 6.2+ [Schwachtgen J.L. et al., 1998]
TRRD-FACTORS	Detailed description of the factors: gene link (**ID**), site link (**AN**), factor name (**TF**), factor origin (**TS**, **TO**), TRANSFAC database link (**NF**), cells (**TC**), and bibliography link (**TR**).	**ID** Hs:VWF **AN** Site: S4238 **TF** Oct-1; **TS** human **NF** TRANSFAC link: T00641 **TO** endogenous **TC** HeLa **TC** HUVEC **TR** [Schwachtgen J.L. et al., 1998]
TRRDBIB	References to the original papers: gene link (**ID**), authors (**AU**), title (**TI**), journal (**SO**), volume (**VL**), issue (**IS**), pages (**PG**), and MEDLINE accession number (**ML**)	**NN** 1579 **ID** Hs:VWF **AU** Schwachtgen J.L., Remacle J.E., Janel N., Brys R., Huylebroeck D., Meyer D., Kerbiriou-Nabias D. **TI** Oct-1 is involved in the transcriptional repression of the von willebrand factor gene promoter. **SO** Blood **VL** 92 **IS** 4 **YR** 1998

TRRD SRS-table	Information content	Entry example
		PG 1247-1258 **ML** MEDLINE:98361770

* Detailed description of the TRRD format was published earlier [Kolchanov et al., 1999].

The SRS system enables to make complex queries and navigation through TRRD and related databases. The links contained in different TRRD database tables may be used for getting an additional information about the sites and regulatory regions.

TRRD database is a constituent part of the GeneExpress System [Kolchanov et al., 1998]. TRRD is linked with other components of the GeneExpress system (http://wwwmgs.bionet.nsc.ru/systems/GeneExpress/) as well as with other databases on molecular biology. The hypertext references introduced into the TRRD documents serve for navigation. The hyperlinks and automated linking by SRS technique provide TRRD integration with the other informational resources and software.

Implementation and results

Current release of the TRRD comprises the description of more than 900 genes and over 4200 transcription factor binding sites. This information was obtained through annotating more than 3000 scientific publications. The dynamics of information volume accretion in the TRRD database is shown in Fig. 1.

Development of TRRD now is directed to description of individual functional gene systems. Among the genes accumulated in TRRD the following functional gene groups are considered: interferon-inducible genes, erythroid-specific regulated genes, genes of lipid metabolism, glucocorticoid-controlled genes, cell cycle-dependent genes, endocrine system genes, heat shock-regulated genes, and plant genes (Table 2). The described representation of information in TRRD helps the user to enter directly the database section of interest and, if needed, to access the other modules of TRRD database.

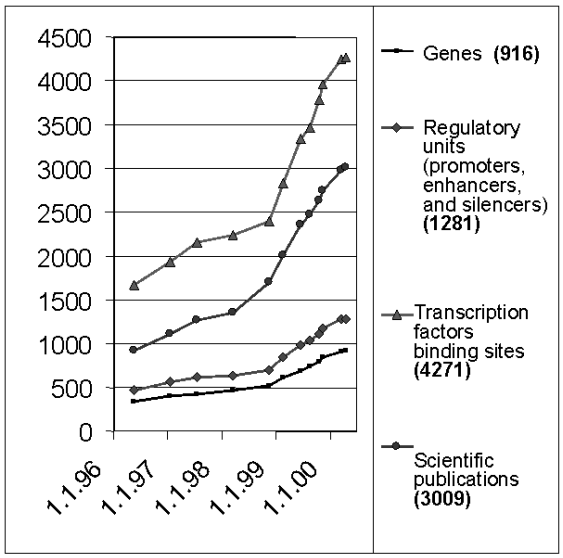


Figure 1. Dynamics of the TRRD database accretion from 1996 to 2000 (dated by April, 15, 2000).

Discussion and conclusions

The TRRD format allows describe gene expression peculiarities in accordance with the developmental stage of an organism, cell differentiation, and cell cycle stage. Besides, organo- and tissue-specificity of gene expression and the influence of external stimuli are considered. It should be emphasized that the expression pattern is related to definite regulatory units and binding sites making a significant impact to realization of the pattern. Thus, TRRD informational model describes the integrity of gene transcription regulation. Due to this fact, the TRRD database may serve as a core for an integration of the other informational resources.

Table 2. Functional sections of the TRRD database.

TRRD section	Genes	Regulatory units	Sites
*Genes of Lipid Metabolism (***LM-TRRD***)*	75	117	455
*Endocrine System Transcription Regulatory Regions Database (***ES-TRRD***)*	89	131	425
Cell Cycle-Dependent Genes (**CYCLE-TRRD**)	51	84	213
Glucocorticoid Controlled Genes (**GR-TRRD**)	52	89	423
*Erythroid-Specific Regulated Genes (***ESRG-TRRD***)*	56	120	505
*Plant Genes (***PLANT-TRRD***)*	140	140	380
*Heat Shock-Induced Genes (***HS-TRRD***)*	80	73	197
*Interferon-Inducible Genes (***IIG-TRRD***)*	98	135	421

As the basal programming software for getting an access to the TRRD database and for its integration with the other informational resources within the frames of the GeneExpress system [Kolchanov et al., 1998], an SRS is applied. The SRS makes possible to use the links between databases for realization of a complex query for solving different biological tasks. Thus, the TRRD database provides a user with good opportunities for data management. It is oriented as to the users dealing with computer analysis (making data samples, the search for principles of gene expression regulation, the testing of hypotheses, etc.) as to researchers engaged in the fields of molecular biology, molecular evolution, cell biology, biotechnology, gene therapy, etc.

For the further development of TRRD, we plan to create a server of applications that provides analysis and processing of the data accumulated in the TRRD and other databases. The server will be accessible both through the CORBA media and standard Common Gateway Interface. For this purpose, besides applying hyperlinks and automated linking between the TRRD tables by the SRS, we plan to make an object-oriented description of informational resources and to use the CORBA technology for getting an access to distributed objects.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grants Nos. 98-04-49479, 98-07-91078, 00-07-90337), Russian Human Genome Program, Ministry of Science and Technology of Russian Federation, Integrated Program of the Siberian Department of the Russian Academy of Sciences. The authors thank T.V. Busygina, V.M. Merkulov, O.E. Belova, S.A. Fedorova, O.G. Smirnova, and V.V. Suslov, for participation in annotating papers for TRRD.

References

1. Bairoch,A. and Apweiler,R., (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.,* 28, 45-48.
2. Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G., Tuli,M.A. (2000) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 28, 19-23.
3. Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., et al. (2000) The Protein Information Resource (PIR). *Nucleic Acids Res.* 28, 41-44.
4. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A., Wheeler,D.L. (2000) GenBank. *Nucleic Acids Res.* 28, 15-18.
5. Etzold,T., Ulyanov,A., Argos,P. (1996) SRS: Information Retrieval System for Molecular Biology Data Banks. Methods in Enzymology, 266, 114.
6. Kel-Margoulis,O.V., Romashchenko,A.G., Kolchanov,N.A., Wingender,E., Kel,A.E. (2000) COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.,* 28, 311-315.
7. Kolchanov,N.A. Ananko,E.A., Podkolodnaya,O.A., Ignatieva,E.V., et al. (1999) Transcription Regulatory Regions Database (TRRD): its status in 1999. *Nucleic Acids Res.,* 27, 303-306.
8. Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., et al. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Res.,* 28, 298-301.
9. Kolchanov,N.A., Ponomarenko,M.P., Kel,A.E., et al. (1998) GeneExpress: a computer system for description, analysis, and recognition of regulatory sequences of the eukaryotic genome. The Sixth International Conference on Intelligent Systems for Molecular Biology June 28 - July 1 1998, Montreal, Canada, pp.95–104.
10. Perier,R.C., Praz,V., Junier,T., Bonnard,C., Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res*. 28, 302-303.
11. Wingender,E., Chen,X., Hehl,R., Karas,H., et al. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.,* 28, 316-319.

DATABASES ON ENDOCRINE SYSTEM GENE EXPRESSION REGULATION: INFORMATIONAL CONTENT AND COMPUTER ANALYSIS

Ignatieva E.V., Busygina T.V., Ananko E.A., Podkolodnaya O.A., Merkulova T.I., Suslov V.V., Pozdnyakov M.A.

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

Keywords: databases, TRRD, GeneNet, gene expression regulation, computer-assisted analysis of regulatory regions

Resume

Motivation:

Systematization and analysis of heterogeneous experimental data on molecular-genetic mechanisms of live systems functioning is an actual problem of up-to-date bioinformatics.

Results:

Databases on endocrine system (ES) gene expression regulation supporting coordination of the most processes of life support in an animal organism has been developed. The informational content of the databases is presented. Structural features of the gene regulatory regions involved in the ES functioning are analyzed.

Availability:

http://wwwmgs.bionet.nsc.ru/mgs/papers/ignatieva/es-trrd/

Introduction

Currently, a rapid accumulation of experimental data on molecular-genetical mechanisms forming the foundation of endocrine regulation of physiological functions is in progress. With this respect, biological databases of various types, GenBank [Benson,D. et al., 2000], EMBL [Baker,W., et al., 2000], SWISS-PROT [Bairoch,A., and Apweiler,R. 2000], provide a possibility for accumulation of information on genes and proteins important for ES functioning. Some particular aspects useful for understanding the mechanisms of endocrine regulation are represented in the databases PATHWAY [Kanehisa,M. and Goto,S. 2000] and RECEPTOR [Nakata,K. et al., 1999]. However, up to now, no efforts were made to unify the whole integrity of data on the functioning of endocrine system at different levels (gene, cell, tissue, organ) within the frames of the single informational system. During the last years, the concept of gene networks is being intensively developed [Kolchanov N.A., in press]. It is based on the idea about coordinated gene expression regulation controlling the definite physiological process. Therefore, by fulfilling the task of theoretical analysis of ES regulation at the level of transcription, we have developed an informational system, on the ground of TRRD, GeneNet, and SAMPLES databases, this system enabling to accumulate experimental data on different levels of ES functioning.

Methods

Information on mechanisms of endocrine functions regulation at the levels of a cell and an organism was accumulated in the GeneNet database [Kolpakov,F.A. et al., 1998], by applying the system for interactive data input through the Internet [Kolpakov,F.A., Ananko,E.A. 1999]. Data on gene transcription regulation were stored in the TRRD database [Kolchanov,N.A. et al., 1999; Kolchanov,N.A., et al., 2000]. The knowledge on the primary sequences of the SF-1 binding sites was classified by using the format of the SAMPLES database [Vorobiev,D.G., 1998]. Recognition of potential transcription factor SF-1 binding sites was produced by means of original method by M Pozdnyakov.

Results

1.ES-TRRD (Endocrine System Transcription Regulatory Regions Database). A section of TRRD database on transcription regulation of the endocrine system genes was developed. This section contains information about 89 genes. Informational content of this section is given in Fig. 1. This section accumulates an information about 131 regulatory regions and 425 transcription factor binding sites. 372 original research publications were annotated to obtain these data. According to the ES-TRRD information, the endocrine system is controlled by

more than 30 various transcription factors. The most significant for ES genes functioning are the factors SF-1, SP-1, GR, Ap-1, Pit1, GATA, COUP-TF, AR, AP2, ER.

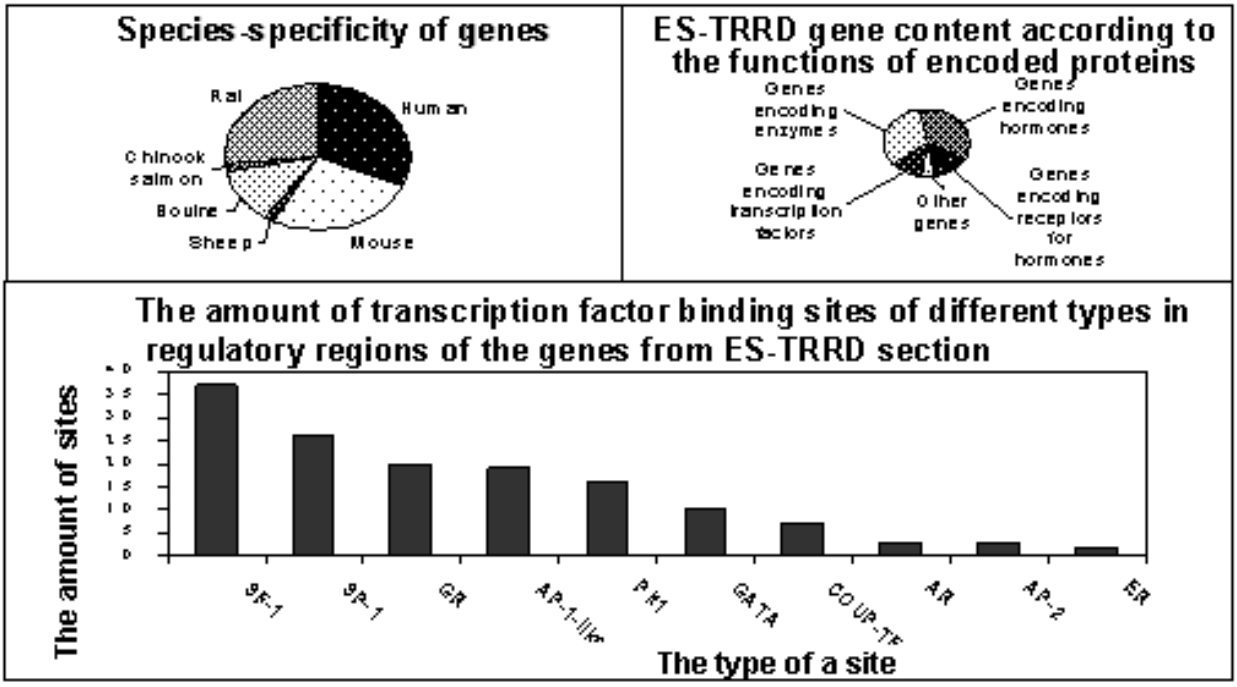


Figure 1. Informational content of the section ES-TRRD.

2.ES-GeneNet database. A section of the GeneNet database containing the information on molecular-genetical mechanisms of endocrine function regulation (Endocrine System GeneNet) was developed. Informational content of this section is given in Table 1. On April, the 1st, 2000, the GeneNet database contained the data on four gene networks, which were represented at four diagrams.

Table 1. Informational content of the diagrams on endocrine regulation of physiological processes from the GeneNet database (by April, 1, 2000).

Name of a diagram*	Physiological process	Number of components		
		Genes	Proteins	Inter-actions
Thyroid system	Regulation of thyroid hormones biosynthesis	9	25	48
Principal cell of CCD	Mechanism of aldosterone impact on the principle cells of CCD in rat kidney	3	18	36
Steroidogenesis (adrenal cortex)	Regulation of glucocorticoids and aldosterone biosynthesis	15	40	80
Steroidogenesis (sex steroides)	Regulation of testosterone and estradiole biosynthesis	12	39	76
Total		39	122	240

3.SAMPLES database. On the basis of information accumulated in the ES-TRRD, we have organized the sample of 5' regulatory regions of the endocrine system (58 entries) and the sample of sequences of SF-1 transcription factor binding sites (34 entries).

4.Analysis. The SF-1 transcription factor binding sites represent the most numerous group among the sites, which are present in regulatory regions of the ES-TRRD (Fig. 1). The binding sites of this transcription factor are present both in regulatory regions of 15 genes of different species, which encode enzymes and some other proteins, directly involved as in biosynthesis of steroid hormones, as in regulatory regions of the other genes of hypothalamic-pituitary-adrenocortical complex together with hypothalamic-pituitary-gonadal complex [Busygina T., this issue]. Our data are in a good accordance to the contemporary knowledge on SF-1 being a key regulator of development and functioning of these two subsystems [Luo X., et al., 1999]. Based on the information accumulated in ES-TRRD, we have analyzed distribution of SF-1 transcription factor binding sites relatively transcription start. It should be noted that 26 out of 34 sites, included into the sample of sequences binding SF-1 factor, are located within the region from -400 to -1 relatively transcription start (see Fig. 2).

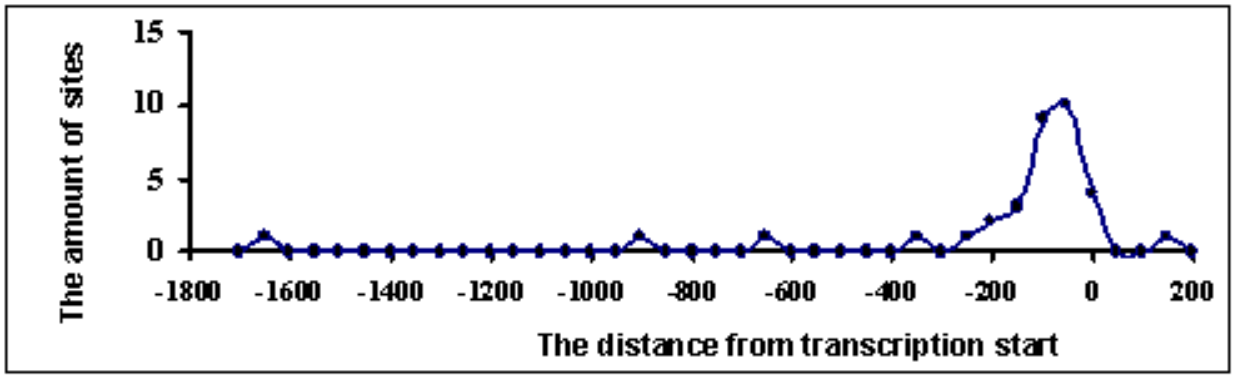


Figure 2. Distribution of SF-1 binding sites in regulatory regions of the genes contained in ES-TRRD.

By using the original computer program for recognition transcription factor binding sites by M. Pozdnyakov, along with the sample of SF-1 binding sites, we have analyzed the properties of regulatory regions of two groups of genes. The first group (denoted as "Yes") includes 18 genes from ES-TRRD, which contain experimentally detected SF-1 transcription factor binding sites, whereas the second group (denoted as "No") contains 23 genes from ES-TRRD, which are not related to the functioning of hypothalamic-pituitary-adrenocortical and hypothalamic-putuitary-gonadal complexes. We have compared the density of potential SF-1 transcription factor binding sites in regulatory regions of the genes entering the groups "Yes" or "No" within the regions in-between -400/-1 and -200/-1. The data shown in Fig. 3 give the evidence that regulatory regions of the genes from the groups "Yes" and "No" differ significantly by the density of potential SF-1 transcription factor binding sites under the level of homology determined so that from 0.1 up to 6 potential binding sites per 100 bp may be detected within the region -200/-1 relatively transcription start in the group "Yes".

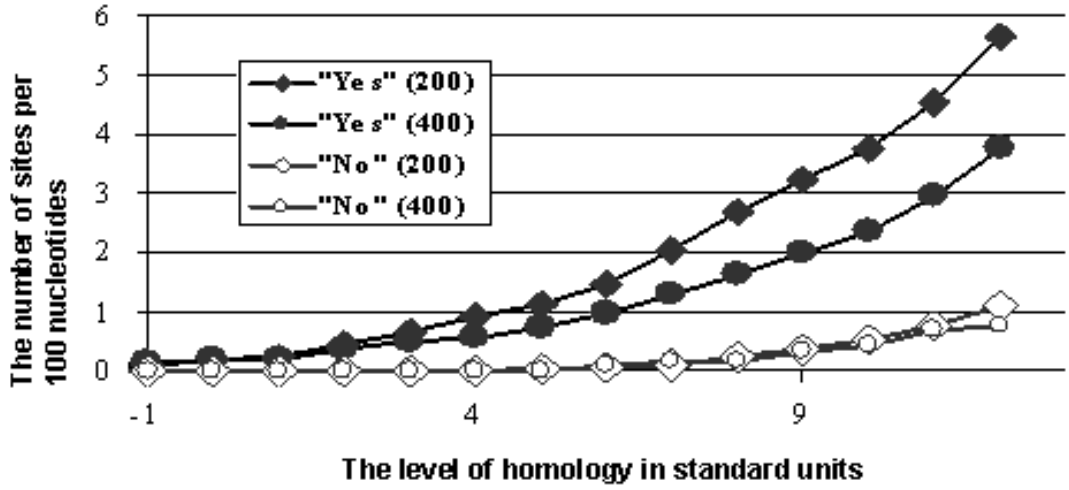


Figure 3. The density of potential SF-1 binding sites in regulatory regions of the genes contained in ES-TRRD. "Yes" (200) and "Yes" (400) - the density of potential SF-1 binding sites in genes from group "Yes" within the regions -200/-1 and -400/-1 respectively; "No" (200) and "No" (400) - the density of potential SF-1 binding sites in genes from group "No" within the regions -200/-1 and -400/-1.

The density of potential binding sites in regulatory regions of genes from the group "Yes" significantly exceeds the analogous value in regulatory regions with the same localization within the genes from the group "No". Notably, the density calculated for the group "Yes" within the region -200/-1 is much more than the same value for the group "Yes" but calculated at the region -400/-1. This result is in a good agreement with the data on distribution of real transcription factor binding sites given in Fig. 2. An attention should be paid to the fact that the densities of potential binding sites for the group of genes "No", revealed by analysis of the sequences within the regions -400/-1 and -200/-1, are practically the same both for the interval of homology considered (Fig. 3) and for the other possible values of this character (data not shown).

Conclusion

We have presented an informational content of developed sections of the TRRD and GeneNet databases on ES genes expression regulation. The data analysis performed at the current stage has revealed the peculiarities of regulation at transcriptional level of the genes participating in regulation of synthesis of steroid hormones and has supported the fruitfulness of an aproach based on the gene network concept. In future, we plan to

supplement the databases by the novel information, this will make easier an analysis of the mechanisms of ES gene network functioning by using various computer methods and approaches.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grants Nos. 98-04-49479, 98-07-91078, 99-07-90203, 00-04-49229, 00-04-49255, 00-07-90337) and the US DOE grant DE-FG02-00ER62893/535228. The authors are grateful to I.V. Lokhova for the help in the work with the literature, to P. Kosarev for the help in extraction of the samples in the format of the SAMPLES datadase, to D.G. Vorobiev – for installation of samples in the Internet, to G.V. Orlova for translation of the paper into English.

References

1. Busygina T. et al., this issue
2. Benson,D., Karsch-Mizrachi,I., Lipman,D., Ostell,J., Rapp,B., Wheeler,D. (2000) GenBank. Nucleic Acids Res., 28, 15-18.
3. Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G., Tuli,M.A. (2000) The EMBL Nucleotide Sequence Database Nucleic Acids Res., 28, 19-23.
4. Bairoch,A., Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000 Nucleic Acids Res., 28, 45-48.
5. Ignat'eva,E.V., Merkulova,T.I., Vishnevsky,O.N., Kel',A.E. (1997) Transcriptional regulation of lipid metabolism genes: description in the TRDD database. Mol. Biol. (Mosk), 31, 684-700
6. Kanehisa,M., Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res., 28, 27-30
7. Kolchanov,N.A., Ananko,E.A., Kolpakov,F.A., Podkolodnaya,O.A., Ignatieva,E.V.,Goryachkovsky,T.N., Stepanenko,I.L. (2000) Gene networks. Russian Journal of molecular Biology, (in press)
8. Kolchanov,N.A., Ananko,E.A., Podkolodnaya,O.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.N., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N., Romashchenko,A.G. (1999) Transcription Regulatory Regions Database (TRRD): its status in 1999. Nucleic Acids Res., 27, 303-306.
9. Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N., Korostishevskaya,I.M., Romashchenko,A.G., Overton,G.C. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. Nucleic Acids Res., , 28, 1, p. 298-301
10. Kolpakov,F.A., Ananko,E.A. Kolesov,G.B. Kolchanov,N.A. (1998) GeneNet: a gene network database and its automated visualization. Bioinformatics, 14, 529 – 537.
11. Kolpakov F.A., Ananko E.A. (1999) Interactive data input into the GeneNet database. Bioinformatics, 15, 713-714
12. Luo,X., Ikeda,Y., Lala,D., Rice,D., Wong,M., Parker,K.L. (1999) Steroidogenic factor 1 (SF-1) is essential for endocrine development and function. J.Steroid.Biochem.Mol.Biol., 69, 13-18.
13. Nakata,K., Takai,T., Kaminuma,T. (1999) . Development of the receptor database (RDB): application to the endocrine disruptor problem. Bioinformatics, 15, 544-552.
14. Vorobiev,D.G., Ponomarenko,J.V., Podkolodnaya,O.A. (1998) Samples And Aligned: Databases For Functional Site Sequences. Proc. I Intern. Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Russia, 1, 58-61.

ASDB: DATABASE OF ALTERNATIVE SPLICING

*[1]Dralyuk I., [1]Brudno M., *[2]Gelfand M.S., [1]Zorn M., [1]Dubchak I.*

[1]National Energy Research Scientific Computing Center, Berkeley, USA
[2]State Scientific Center GosNIIGenetika, Moscow, Russia
e-mail: misha@imb.imb.ac.ru
*Corresponding author

Keywords: database, alternative splicing, gene expression

Resume

Alternative splicing is an important mechanism of gene expression. It has been estimated that at least one third oh human genes are alternatively spliced (Mironov and Gelfand, 1998; Hanke et al., 1999; Mironov et al., 1999). However, only a few compilations of alternatively spliced genes has been published, and these compilations are very limited: they cover alternative splicing in individual tissues (Stamm et al., 1994) or species (Kent and Zahler, 2000), or alternative splicing in specific gene families (Atamas, 1997; Lopez, 1995). Only the nematode alternative splicing database (Kent and Zahler, 2000) is available via the Internet.

Recently developed version 2 of ASDB (Alternative Splicing Data Base) consists of two divisions: proteins and genomic sequences (Dralyuk et al., 2000). It is available at http://www.cbcg.nersc.gov/asdb.

The protein division contains 1922 protein sequences, as compared to 1663 entries in version 1 (Gelfand et al., 1999). It was developed by selecting SwissProt (Bairoch and Apweiler, 1999) entries using search for the words `alternative splicing' (usually in the CC lines) and `varsplic' (in the FT lines). Some entries describe just one alternatively spliced variant, some (those that include the `varsplic' field) indicate several variants.

In order to group proteins from different entries that could arise by alternative splicing of the same gene, we developed the clustering procedure (Gelfand et al., 1999). Two proteins were linked if they had a common fragment of at least 20 amino acids, and clusters were initially defined as maximum connected groups of linked proteins. Each cluster was represented by multiple alignment of its members constructed using CLUSTALW (Thompson et al., 1997) (Fig. 1). After version 1 of ASDB was published, it turned out that some clusters were chimeric, in the sense that they contained members of multigene families, but not alternatively spliced variants of one gene. Therefore the multiple alignments were subject to additional automated and manual analysis aimed at removal of chimeric clusters.

```
2ACA_HUMAN      LVDLEPKSKV SSPIEKVSPS CLTRIIETNG HKIEEEDRAL LLRILESIED
2ACB_HUMAN      .......... .......... .......... .......... ..........

2ACA_HUMAN      FAQELVECKS SRGSLSQEKE MMQILQETLT TSSQANLSVC RSPVGDKAKD
2ACB_HUMAN      .......... .......... .MMIKETSLR RDPDLRGELA FLARGCDFVL

2ACA_HUMAN      TTSAVLIQQT PEVIKIQNKP EKKPGTPLPP PATSPSSPRP LSPVPHVNNV
2ACB_HUMAN      PSRFKKRLKS FQQTQIQNKP EKKPGTPLPP PATSPSSPRP LSPVPHVNNV

2ACA_HUMAN      VNAPLSINIP RFYFPEGLPD TCSNHEQTLS RIETAFMDIE EQKADIYEMG
2ACB_HUMAN      VNAPLSINIP RFYFPEGLPD TCSNHEQTLS RIETAFMDIE EQKADIYEMG

2ACA_HUMAN      KIAKVCGCPL YWKAPMFRAA GGEKTGFVTA QSFIAMWRKL LNNHHDDASK
2ACB_HUMAN      KIAKVCGCPL YWKAPMFRAA GGEKTGFVTA QSFIAMWRKL LNNHHDDASK
```

Figure 1. A typical ASDB protein cluster. Dotted line: common segments; no line: alternative segments; double line: spurious matches.

The protein entries are marked with different symbols to allow for easy differentiation among the three types of entries: proteins that are part of the ASDB clusters and the corresponding multi-alignments, proteins that have information on different variants in the associated SWISS-PROT entries (in the "varsplic" fields), and proteins for which information on the variants is not available at the present time. ASDB hass internal links between entries and/or clusters, as well as external links to MEDLINE, GenBank and SWISS-PROT entries.

The DNA division contains 2486 entries (Fig. 2). It was generated by collecting all GenBank (Benson et al., 1999) entries containing the words `alternative splicing' and further selection of those entries that contain complete gene sequences (all CDS fields are complete, i.e., they do not have continuation signs).

ASDB can be searched using MEDLINE, SWISS-PROT and GenBank identifiers and accession numbers. Standard context search can be performed over SWISS-PROT and GenBank keywords, description, taxonomy, comment fields and feature tables. Boolean logic queries are supported to allow for a wider range of queries.

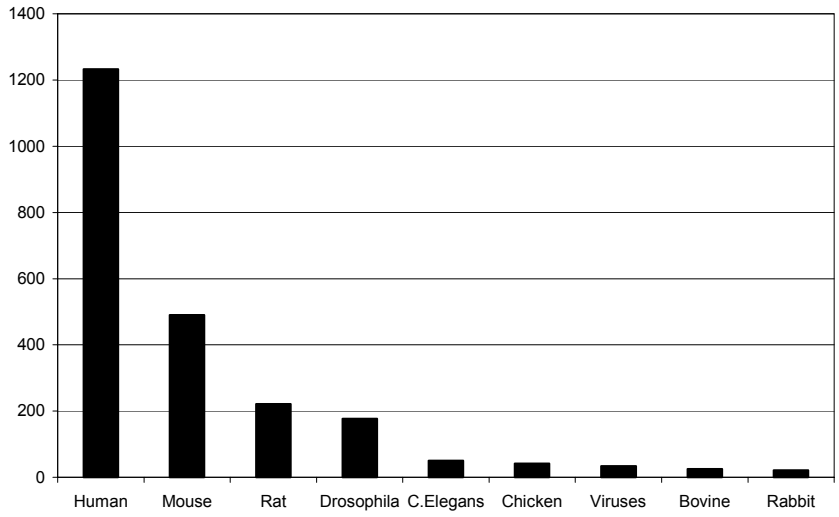


Figure 2. The number of ASDB nucleotide entries for the most represented species.

Acknowledgements

We are grateful to Andrey Mironov for useful discussions. This study was supported the Director, Office of Energy Research, Office of Biological and Environmental Research, of the US Department of Energy under Contract No. DE-ACO3-76SF00098 and by grants from the Merck Genome Research Institute (244), the Russian Fund of Basic Research (99-04-48247 and 00-15-99362), the Russian State Scientific Program "Human Genome", and INTAS (99-1476).

References

1. Atamas, S.P. (1997) *Life Sci.,* 61, 1105-1112.
2. Bairoch, A. and Apweiler, R. (1999) *Nucleic Acids Res.,* 27, 49-54.
3. Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, B.F.F., Rapp, B.A. and Wheeler, D.L. (1999) *Nucleic Acids Res.,* 27, 12-17.
4. Dralyuk, I., Brudno, M., Gelfand, M.S., Zorn, M. and Dubchak, I. (2000) ASDB: database of alternatively spliced genes. *Nucleic Acids Res.,* 28, 296-297.
5. Gelfand, M.S., Dubchak, I., Dralyuk, I. and Zorn, M. (1999) ASDB: database of alternatively spliced genes. *Nucleic Acids Res.,* 27, 301-302.
6. Hanke, J., Brett, D., Zastrow, I., Aydin, A., Delbruck, S., Lehmann, G., Reich, J. and Bork, P. (1999) *Trends Genet.,* 15, 389–390.
7. Kent, W.J. and Zahler, A.M. (2000) *Nucleic Acids Res.,* 28, 91-93.
8. Lopez, A.J. (1995) *Dev. Biol.,*, 172, 396-411.
9. Mironov, A.A. and Gelfand, M.S. (1998) Gene recognition using EST data: unexpectedly frequent alternative splicing of human genes. *1st Conf. BGRS-98,* vol. 2, pp. 249-250.
10. Mironov, A.A., Fickett, J.W. and Gelfand, M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.,* 9, 1288-1293.
11. Stamm, S., Zhang, M.Q., Marr, T.G. and Helfman, D.M. (1994) *Nucleic Acids Res.,* 22, 1515-1526
12. Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. (1997) *Nucleic Acids Res.,* 25, 4876-4882.

CYCLE-TRRD: A DATABASE ON TRANSCRIPTIONAL REGULATION OF CELL CYCLE-DEPENDENT GENES

Kel-Margoulis O.V., Kel A.E.

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

Keywords: transcriptional regulation, cell cycle, E2F/DP family, database, gene regulatory regions, transcription factors, binding sites

Resume

Motivation:

Study of transcriptional regulation mechanisms of eukaryotic genes is one of the key problems of molecular genetics. In particular, systematic collection of information on transcriptional regulation of genes controlling cell cycle progression is essential for studding mechanisms of the normal cell proliferation and differentiation as well as possible reasons of uncontrolled proliferation and tumor development.

Results:

A CYCLE-TRRD database is developed that contains information on transcriptional regulation of functionally interrelated group of genes controlling cell cycle (55 genes). A set of binding sites for transcription factors of E2F family – one of the major regulators of cell cycle, is constructed on the base of information collected in the CYCLE-TRRD database (43 sites in 31 gene).

Availability:

The database CYCLE-TRRD is availiable at the server of the Institute of Cytology and Genetics SB RAS (http://wwwmgs.bionet.nsc.ru/mgs/papers/kel_ov/celcyc/).

Introduction

The CYCLE-TRRD is a section of the TRRD database that has been developing in the Laboratory of theoretical genetics of the Institute of cytology and genetics SB RAS since 1993 (Kolchanov et al., 2000). Previous release of the CYCLE-TRRD database was described earlier (Kel O.V. and Kel A.E., 1997).

Cell cycle is carried out through subsequent modifications of chromatin structure in a cell until the cell divides into two daughter cells. The central event of the cell cycle is DNA synthesis. During the pre-synthetic period the cell prepares itself to the replication. This period is called G1 phase. Period of replication is called S-phase (synthetic). For DNA replication a rather wide range of enzymes and co-factors is required. Nucleosome packaging requires *de novo* synthesis of histones. Expression of the genes encoding these proteins reaches maximum during S-phase. After replication is completed and genetic material is reduplicated the cell enters the post-synthetic period (G2 phase). During G2 phase the cell prepares itself for the division into two daughter cells (Mitosis, M-phase). Two critical checkpoints have been identified: G1/S and G2/M.

Currently the central role of E2F family factors in the regulation of genes whose expression depends on the cell cycle stage is commonly accepted (Farnham et al., 1993; DeGregori et al., 1995 and others). First, factors of the E2F family take part in transcription regulation of genes that are essential for DNA replication and formation of nucleosome structure. Second, E2F family members form complexes with other regulatory proteins to control transcription of a particular set of genes thus coordinate S-phase entry. E2F target genes include: 1) a number of genes for the components of cell cycle machinery (cyclins and cyclin-depending kinases, the *e2f-1* and *e2f-2* genes, a gene for a tumor suppressor protein pRB); 2) genes for transcription factors involved in diverse intracellular processes (families Myc and Myb); 3) genes for enzymes and other proteins that are necessary for DNA replication; 4) some genes providing DNA reparation; 5) genes encoding chromatin components (histones).

Numerous data suggest that E2F factors are essential for S-phase specific gene transcription as well as for the inhibition of some genes in G0 and early G1 phases (Johnson et al., 1994; Neuman et al., 1994; Luo et al., 1998; Magnaghi-Jaulin et al., 1998). Levels and activity of E2F factors reflect the integrative response to all proliferative and/or antiproliferative signals accepted by the cell.

We developed the CYCLE-TRRD database that contains information on transcriptional regulation of functionally interrelated group of genes controlling cell cycle. A set of binding sites for transcription factors of E2F family is constructed on the base of information collected in the CYCLE-TRRD database.

Content of the CYCLE-TRRD database

Current release of the CYCLE-TRRD database contains information on transcriptional regulation of 55 vertebrate genes (Table 1). Modular hierarchical organization of gene transcriptional regulatory regions is presented in the database. Gene expression patterns (dependence on the cell cycle phase, on the type and differentiation status of the cells, on the induction by external signals) are also presented in the database.

Table 1. Genes collected in the CYCLE-TRRD database.

Gene classification	Gene name	Acc.number of genes in the CYCLE-TRRD
Transcription factors of E2F family	e2f-1 Hs, Mm, Cc; e2f-2 Hs	A00347, A00348, A00655, A00701
Retinoblastoma family	RB1, Hs; p107, Hs	A00329, A00290
Cyclins	cycD1, Hs; cycD2, Hs; cycD3, Hs; cycE Hs, Mm; cycA Hs, Mm cycB1 , Hs; cycG , Mm	A00352, A00353, A00360, A00365, A00366, A00657, A00700, A00711, A00712
Cyclin-dependent kinases	cdc2 Hs, Rn; cdc7, Mm	A00355, A00356, A00658
Phosphatases	cdc25A, cdc25C (человек)	A00354, A00779
Inhibitors of the cyclin-dependent kinases	p21waf1, Hs; p15ink4b, Hs; p16ink4a, Hs	A00359, A00893, A00894
p53 and functionally related genes	p53 Hs, Mm; mdm2 , Mm; p14ARF, Hs	A00307, A00308, A00438, A00656
TF of AP-1 family	c-fos, Hs, Mm; fosB, junB, Mm; c-jun, Hs	A00028, A00095, A00107, A00110, A00182
TF of Myc and Myb families	c-myc , Hs, Mm; N-myc, Mm; B-myb , Mm	A00191, A00108, A00371,A00441
Genes involved in DNA replication and reparation	cdc6, Hs; orc1, Hs; DNA pol.α , Hs; h2a.1, Hs; UDG, Hs; DNA pol.β, Mm; h2a.x, Mm; htf9a , Mm; dhfr, Hs, Mm, Cs; tk, Hs, Mm; PCNA, Hs, Mm; cad, Ma.	A00669, A00710, A00370, A00437, A00024, A00063, A00023, A00298, A00367, A00287, A00446, A00445, A00368, A00369, A00447, A00448
Viral genes regulated by E2F family	Ad 2 EIIaE1, Ad 5 E1A, EBNA1	A00788, A00787, A00444

Cc - Coturnix coturnix (quail); Cs - Cricetulus sp. (Chinese hamster); Hs - Homo sapiens; Ma - Mesocricetus auratus (golden hamster); Mm - Mus musculus; Rn - Rattus norvegicus

Set of the binding sites for E2F transcription factor family.

Based on the information collected in the CYCLE-TRRD database a set of E2F binding sites is constructed that contains 43 sites within 31 different genes (Table 2). We described general characteristics of these sites and of the promoters of corresponding genes.

Table 2. E2F binding sites in transcriptional regulatory regions of genes presented in the CYCLE-TRRD database. TATA-box presence in the promoters of E2F-dependent genes is shown. Positions correspond to the sequence shown by capital letters.

N[1)]	CYCLE-TRRD[2)]	Gene name	TATA-box	Position of E2F sites relatively to transcription start	Sequence of E2F sites
1	S1835	c-myc, Hs	+	-69 to -58	gCTTGGCGGGAAAaaga
2	S1836			-42 to -31	gGATCGCGCTGAGtata
3	S398	c-myc, Mm	+	-81 to –69	gCTTGGCGGGAAAaaga
4	S1926	N-myc, Mm	+	-142 to –131	tTTTGGCGCGAAAggct
5	S1927			-127 to –116	cTTTGGCGCCTCCcctg
6*	S2353	B-myb, Mm	-	-212 to –201	aCTTGGCGGGAGAtagg
7	S1791	e2f-1, Hs	-	-31 to -20	cTTTCGCGGCAAAaagg
8	S1792			-14 to -3	cTTTGGCGCGTAAaagg
9	S1786	e2f-1, Mm	-	-40 to -29	cTTTCGCGGCAAAaagg
10	S1787			-23 to –12	aTTTGGCGCGTAAaagt
11	S3143	e2f-1, Cc	-	-16 to –5	cTTTCGCGGCAAAaagg
12	S3144			+2 to +13	aTTTGGCGCGCAAaggc
13	S1590	RB1, Hs	-	+91 to +102	tTTTCCCGCGGTTggac
14	S1343	p107, Hs	-	-16 to –5	tTTTCGCGCGCTTtggc
15	S1345			-6 to +6	cTTTGGCGCAGGTggtt
16	S1858	cyclin D1, Hs	-	-53 to –42	gTTTGGCGCCCGCgccc
17	S1909	cyclin E, Hs	-	-16 to -5	gTTCCGCGCGCAGggat
18	S1910			+7 to +18	aTGTCCCGCTCTGagcc
19	S3256	cyclin E, Mm	-	-24 to –7	gGCGGGCGCGAGGgcgg
20	S1853	cyclin A, Hs	-	at –37	tAGTCGCGGGATActtg
21	S1870	cdc2, Hs	-	-131 to –114	cTTTCGCGCTCTAgcca
22	S1874			-27 to –13	cTTTAGCGCGGTGagtt
23	S1879	cdc2, Rn	-	-130 to –119	cTTTCGCGCTCTGcact
24	S3147	cdc6, Hs	-	-43 to –32	cTTTGGCGGGAGGtggg
25	S3148			-9 to +6	aTTTGGCGCGAGCgcgg
26	S3257	Orc1, Hs	-	-10 to +2	gATTGGCGCGAAGtttt
27	S1923	DNA pol α, Hs	-	-139 to –128	aCAGGGCGCCAAAcgcg
28	S2260	dhfr, Hs	-	-10 to +2	aTTTCGCGCCAAActtg

N[1)]	CYCLE-TRRD[2)]	Gene name	TATA-box	Position of E2F sites relatively to transcription start	Sequence of E2F sites
29	S2365	dhfr, Mm	-	-10 to +2	aTTTCGCGCCAAActtg
30*	S88	dhfr, Cs	-	-62 to –51	aTTTCGCGCCAAActtg
31	S1914	tk, Mm	-	-80 to –69	aGTTCGCGGGCAAatgc
32	S1338	cad, Ma	-	+74 to +85	cTTCGGCGCGCGGtgtt
33	S2334	PCNA p120, Hs	-	в 1-ом интроне	tTTTCGCGCCAAAgtca
34	S2381	htf9a/RanBP-1, Mm	-	-115 to –104	tTTTGGCGGGAAGcgcg
35	S2288	histone H2A.1, Hs	+	-50 to –39	tTTTCGCGCCCAGcagc
36	S2378	histone H2A.X, Mm	+	-256 to –245	aTTTCGCGCGCTCtaca
37	S3495	UDG, Hs	-	-112 to -101	tTTTGCCGCGAAAagac
38	S3673	Ad type 2 EIIaE1	-	-68 to -52	tTTTCGCGCTTAAattt
39	S3674			-48 to –32	aAAGGGCGCGAAActag
40	S3676	Ad type 5 E1A	+	-288 to -272	tTTTCGCGCGGTTttag
41	S3677			-225 to -209	tTTTCGCGGGAAAactg
42	S2185	Epstein-Barr Virus, EBNA1	+	+191 to +204	aAGGCGCGGGATAgcgt
43	S2187			+218 to +234	aGATGGCGGGTAAtaca

1) Site number in this Table ;
2) Site acc. number in the CYCLE-TRRD database (*http://srs5.bionet.nsc.ru/srs5/*);
* - position are given relatively to the start of translation; Species abbreviation is the same as in the Table 1.

Few E2F-dependent genes contain TATA box. They are proto-oncogenes of Myc family, histone genes and two viral genes (these genes are marked by "+" in the Table 2). The majority of genes presented in the Table 2 (24 out of the 31) do not contain TATA box (marked by "-"). They include genes encoding components of the cell cycle machinery; enzymes and co-factors of DNA replication and reparation; and gene for transcription factor B-myb.

In many cases E2F sites are located within the first 150 bp upstream of the transcription start site. In promoters of *p107*, *cdc6*, *orc1* and *dhfr* genes E2F sites overlap the start site (Table 2, NN 15, 25, 26, 28-30). In some cases E2F sites are situated in the non-coding region of the 1st exon (NN 12,13,18,32) or in the 1st introne (N 33).

As one can see in the Table 2, E2F-dependent genes often have two binding sites for the E2F family factors. These genes are: *e2f-1* of human, mouse and quail (NN 7-12); proto-oncogenes *c-myc* and *N-myc* (NN 1,2,4,5); *p107* (NN 14,15); cyclin E (NN 17,18); EBNA1-nuclear antigen of the Epstein-Barr virus EBNA1and EIIaE1 – adenoviral gene (NN 38,39,42,43). In all these cases E2F sites are situated in close proximity to each other: immediately adjacent or separated by 10-15 bp only. However, in the promoters of some genes E2F sites are separated by rather long distance. For instance, in the *cdc6* promoter E2F sites are separated by 30 bp (Table 2, NN 24,25), in the promoter of the adenoviral gene E1A – by 50 bp (NN 40,41), and in the human *cdc2* promoter – by 85 bp (NN 21,22).

Information comprised in the CYCLE-TRRD database suggests that E2F binding sites are the most frequent cis-elements in the structure of transcriptional regulatory regions of cell cycle-dependent genes. Characteristic features of the structure and distribution of E2F binding sites could be used for the development of computer methods of potential site recognition within newly sequenced regulatory sequences.

References

1. DeGregori J., Kowalik T., and Nevis J.R. (1995) *Mol. Cell. Biol*., 15, 4215-4224.
2. Farnham P.J., Slansky J.E., and Kollmar R. (1993) *Biochim. Biophys. Acta*, 1155, 125-131.
3. Hagemeier C., Cook A., and Kouzarides T. (1993) *Nucleic Acids Res*., 21, 4998-5004.
4. Johnson D.G., Ohtani K., and Nevis J.R. (1994) *Genes Dev*., 8, 1514-1525.
5. Kel O.V. and Kel A.E.. (1997) *Mol. Biol.(Mosk.)*, 31, 548-561.
6. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., and Overton G.C. (2000) *Nucleic Acids Res*., 28, 298-301.
7. Luo R.X., Postigo A.A., and Dean D.C. (1998) *Cell*, 92, 463-473.
8. Magnaghi-Jaulin L., Groisman R., Naguibneva I., Robin P., Lorain S., Le Villain J.P., Troalen F., Trouche D., and Harel-Bellan A. (1998) *Nature*, 391, 601-605.
9. Neuman E., Flemington E.K., Sellers W.R., and Kaelin W.G. (1998) *Mol. Cell. Biol*., 14, 6607-6615.

LOCUS CONTROL REGIONS: DESCRIPTION IN A DATABASE

Podkolodnaya O.A. and Levitsky V.G.

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

Keywords: transcription regulation, regulatory gene regions, LCR, databases

Resume

A database was created on the structure-functional organization of the locus-control regions (LCRs) in eukaryotic genomes. The database contains two sections including (i) formalized descriptions of LCRs and (ii) hypertextual description of the LCR structure and function. Both sections are linked by hypertexts between each other and with external databases such as TRRD, SWISS-PROT, EMBL, and Medline.

Availability:

The database is maintained under SRS and is available by the Internet via
http://wwwmgs.bionet.nsc.ru/mgs/dbases/lcr/

Introduction

In the middle of 80-ties, a regulatory element of the novel type was detected that provides coordinated stage- and tissue-specific regulation of human β-globin locus genes transcription [Grosveld et al., 1987]. This element was named as the Locus Control Region (LCR). Later on, analogous regulatory regions were found in the other gene loci.

Some LCRs may differ significantly by the content of the constituent elements; by the content of genes controlled by these LCRs; and by the LCR disposition respectively regulated genes. LCR may regulate both expression of the whole gene cluster (e.g., human β-globin) and the separate gene (e.g., human adenosine deaminase). LCRs regulating the clusters of α- and β-globin genes in mammals, chicken lysozyme, mouse glycophorin are located in the 5' region. On the contrary, LCR regulating growth hormone and human CD2 genes are located in the 3' region, whereas for adenosine deaminase gene, the corresponding regulatory region is situated in the first intron. Although it was found that LCR might regulate not only the whole gene cluster but also the separate gene, the traditional name of this regulatory element is not changed. Formally, as a locus controlling region (LCR) we denote the DNA fragment (or the grouped DNA fragments), such that in transgene experiments it provides the high level of tissue-specific expression of the linked gene integrated into transgene construction, proportionally the copy number of this construction and independently the place of its insertion in genome [Li et al., 1999; Grosveld, 1999].

LCRs provide coordinated tissue- and stage-specific expression of the genes entering the regulated cluster. Currently, the data are available on more than 30 LCRs and LCR-like elements. As a rule, the LCR structure is complex. Its particular elements are marked by sites with the high sensitivity to DNAase I (HSS), which could be tissue-specific or expressed in all the tissues [Grosveld et al., 1987; Chung et al., 1993]. Functionally, the elements entering different LCRs are represented by (a) boundary elements or insulators [Abruzzo and Reitman 1994; Jackson et al., 1996], (b) the chromatin domain opening elements [Ortiz et al., 1997; Festenstein et al., 1996], (c) facilitator elements [Aronow et al., 1995], and (d) enhancers. Enhancers are found almost in all the LCRs. Generally, the most part of the LCR enhancer activity is produced by the core sequences with the length of about 200 bp, where transcription factor binding sites are located. A special type of enhancers, in addition to the classic ones, was found among LCRs. Activity of these enhancers depends upon their orientation and is revealed during their integration into the chromatin, but is not detected in the experiments with the transient transfection [Terajima et al., 1995].

Each of the elements listed above makes its own impact into transcription regulation of the corresponding gene cluster, but only interaction of the whole integrity of elements provides the proper LCR functioning.

Results and discussion

In order to describe structure-functional LCR features, we have developed a special format and created a specialized database LCR-TRRD. Informational fields of this database produce the knowledge on the structure of the gene locus under regulation; the structure of the LCR itself; its individual components; along with information on experimental evidence supporting these facts. The database is linked by the hypertexts to the

table of genes (TRRDGENES) in the TRRD database [Kolchanov et al., 2000], and to the databases EMBL, SWISS-PROT, and MEDLINE.

In Figure 1, the schemes of some LCRs are given together with the gene loci regulated by them and described in the LCR-TRRD database.

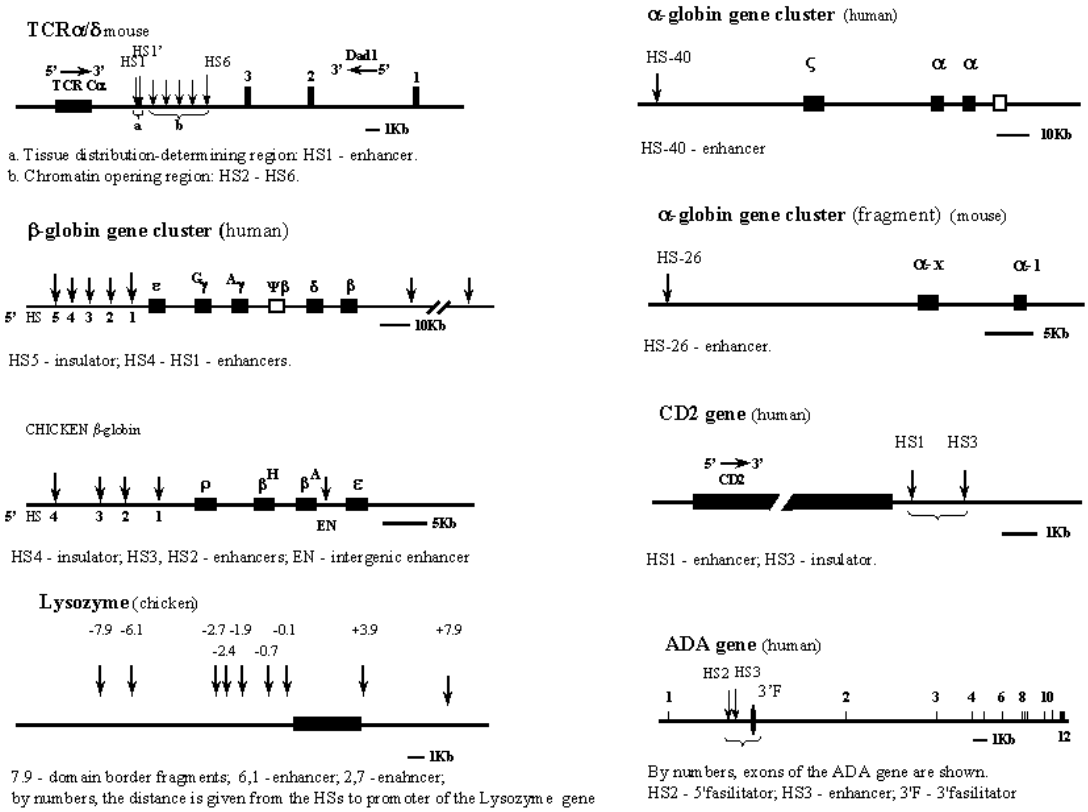


Figure 1. The schemes of some LCRs with the gene loci regulated by them.

Each LCR is supplied by an enhancer as a mandatory element. It was proved that enhancers in the LCRs are tissue-specific, as a rule, and, hence, exactly they are responsible for the LCR's tissue-specificity in the most cases. Tissue-specificity of enhancers themselves is produced by the presence of the tissue-specific transcription factors. So, for the erythroid-specific LCRs, the sites binding the factors NF-E2/AP1 and GATA-1 are typical. In enhancer of the lysozyme gene cluster LCR, the binding site for the mieloid-specific factor PU.1 is located, whereas within the enhancer of the growth hormone gene, the pituitary-specific factor Pit-1 was found. The LCR-regulated transcription in T-lymphocytes (T-cell receptor α/δ, adenosine deaminase, CD2) is marked by the presence of SOX4 and LEF-1/TCF-1 transcription factors binding sites. It is necessary to note that both these factors are specifically expressed in T-lymphocytes, being referred to the HMG protein family, which bind to DNA and cause its bending. In the LCR-TRRD database, there are links to the other sections of TRRD database containing detailed description of transcription factor binding sites found in the corresponding LCR elements.

Enhancers are the best studied LCR elements, but they are not the only ones. The principles of organization and functioning of the other regulatory structures are not completely studied yet. It is supposed that the LCR's action is determined by its ability to establish and maintain the open chromatin domain [Dillon and Grosveld, 1993]. Relying on the fact that nucleosome organization of DNA is laid in the basic level of its packaging, we have performed the analysis and search in the locus control DNA sequences of the features responsible for the nucleosome binding site formation. For the nucleosome site recognition, we have used the computer program [Levitsky and Katokhin, 2000], based on the discriminant analysis method. If the recognition function value tends to +1, this fact evidences about the better ability of the sequence analyzed to nucleosome positioning. In Figure 2, the results of such analysis of the DNA sequence of the chicken β-globin gene cluster LCR are shown. It can be seen that in the HSS2 and HSS1 containing DNA regions, the deviation of the recognition function values from the basal level is detected, this pointing out to decrease of nucleosome potential in these regions.

The HSSs shown in this Figure are the typical tissue-specific hypersensitive sites. They can be found in all the erythroid cell lines and are absent in the non-erythroid ones. Usually, hypersensitive to DNAase I sites are associated with the active regulatory regions that lack canonical nucleosomes. In this case, a nucleosome may be either missing or be partially destroyed. Probably, the contextual DNA features of these regulatory regions

may be responsible for decrease in nucleosome binding potential, thus, favoring to tissue-specific pattern of transcription regulation.

Acknowledgments

The work was partially supported by the Russian Human Genome Program, Ministry of Science and Technology of Russian Federation, Russian Foundation for Basic Research (grants Nos. 98-04-49479, 98-07-91078, 99-07-90203, 00-07-90337, 00-04-49229). The authors are grateful to Belova O.E. for the database filling and helpful discussions; to N.L.Podkolodny, and D.A. Grigorovich for SRS and software installment and support; and to G.V. Orlova for translation of the paper into English.

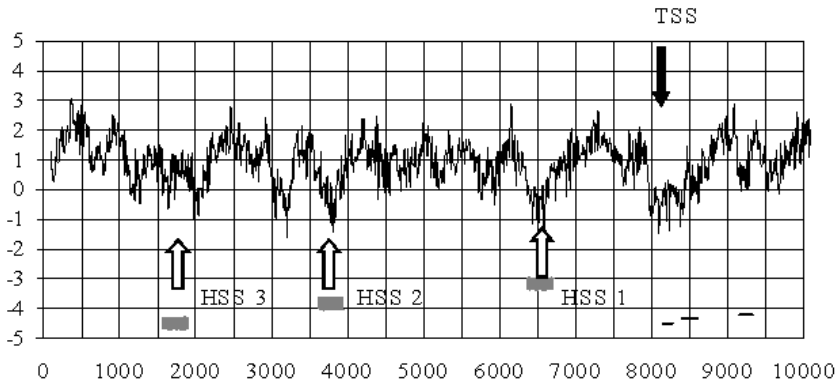


Figure 2. The profile of the function for recognition of nucleosomal sites for the LCR of the β-globin locus in chicken (EMBL AC L17432). The locations of hypersensitive sites and exons are marked below. Positions of hypersensitive sites (HSS) and transcription start sites (TSS) are marked by arrows.

References

1. Abruzzo L.V., Reitman M. (1994) Enhancer activity of upstream hypersensitive site 2 of the chicken beta-globin cluster is mediated by GATA sites. *J.Biol.Chem.*, 269, 32562-32571.
2. Aronow B.J., Ebert C.A., Valerius M.T., Potter S.S., Wiginton D.A., Witte D.P., Hutton J.J. (1995) Dissecting a locus control region: facilitation of enhancer function by extended enhancer-flanking sequences.//*Mol. Cell.Biol.* 15, 1123-1135.
3. Chung J.H., Whiteley M., Felsenfeld G. (1993) A 5' element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. *Cell*, 74, 505-514.
4. Dillon N., Grosveld F. (1993) Transcriptional regulation of multigene loci: multilevel control. *Trends Genet.* 9, 134-137
5. Festenstein R., Tolaini M., Corbella P., Mamalaki C., Parrington J., Fox M., et al. (1996) Locus control region function and heterochromatin-induced position effect variegation. *Science*, 271, 1123-1125.
6. Grosveld F., van Assendelft G.B., Greaves D.R., Kollias G. (1987)Position-independent, high-level expression of the human beta-globin gene in transgenic mice.*Cell.* ,51, 975-985.
7. Grosveld F. (1999) Activation by locus control regions? *Curr. Opin. Genet. Dev.*, 9,152-157.
8. Jackson J.D., Petrykowska H., Philipsen S., Miller W., Hardison R. (1996) Role of DNA sequences outside the cores of DNase hypersensitive sites (HSs) in functions of the beta-globin locus control region. Domain opening and synergism between HS2 and HS3. *J.Biol.Chem.*, 271, 11871-1188.
9. Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., et al. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Res.,* **28**, 298-301.
10. Levitsky V.G., Katokhin A.V. Inherent modular promoter structure and its application for recognition tools development. *Computational technologies* 2000, (in press).
11. Li Q., Harju S., Peterson K.R.(1999) Locus control regions: coming of age at a decade plus. *Trends Genet.*, 15, 403-408.
12. Ortiz B.D., Cado D., Chen V., Diaz P.W., Winoto A.(1997) Adjacent DNA elements dominantly restrict the ubiquitous activity of a novel chromatin opening region to specific tissues. *EMBO J.*, 16, 5037-5045.
13. Terajima M., Nemoto Y., Obinata M. (1995) Inducible expression of erythroid-specific mouse glycophorin gene is regulated by proximal elements and locus control region-like sequence. *J Biochem (Tokyo)*, 118, 593-600

REPRESENTATION OF INFORMATION ON ERYTHROID GENE EXPRESSION REGULATION IN THE GENEEXPRESS SYSTEM

Podkolodnaya O.A. , Stepanenko I.L., Ananko E.A., Vorobiev D.G.

Institute of Cytology and Genetics, Novosibirsk, Russia
opodkol@bionet.nsc.ru
*Corresponding author

Keywords: transcription regulation, gene expression regulation, erythroid-specific genes, gene networks, regulatory gene regions, databases

Resume

Within the frames of the GeneExpress system, an informational resource devoted to erythroid-specific gene expression regulation was developed. This resource contains information about structure-functional organization of regulatory regions of genes specifically regulated in erythroid cells; patterns of their expression; samples of transcription factor binding sites regulating expression of these genes; and the gene network, which acts in erythroid cell in the process of its differentiation under the action of erythropoietin.

Results and discussion

The GeneExpress system enables to represent and analyze various data on expression regulation of genes in eukaryotes [Kolchanov et al., 1999a]. Within the frames of this system, a section was developed, which is devoted to regulation of erythroid-specific gene expression. The main bulk of information contained in this section is distributed between 3 modules of the GeneExpress system, the databases ESRG-TRRD, SAMPLES, and GeneNet system.

In ESRG-TRRD database, an information is accumulated about structure-functional organization of regulatory regions within erythroid-specific genes, along with expression patterns of these genes (http://wwwmgs.bionet.nsc.ru/mgs/papers/podkolodnaya/esg-trrd/). Information of this section is represented in the format of the TRRD database [Kolchanov et al., 1999b]. At present, the ESRG-TRRD database stores the data on more than 500 transcription factor binding sites, which are organized in 120 regulatory units of 56 erythroid-specific genes (Table 1). Besides, the database contains the description of 242 expression patterns of these genes.

Table 1. Content of the ESRG-TRRD database.

Group	Number of entries	Number of sites	Number of regulatory units
Globins	19	222	43
Enzymes of the heme biosynthesis pathway	7	50	16
Other enzymes	5	59	11
Transcription factors	9	43	16
Cellular surface antigens and others	13	86	28
Globin locus-controlling regions	3	45	6

It was noted earlier that transcription factor GATA1 is a key regulator of erythroid-specific genes transcription. Regulatory regions of all the genes, transcription of which is specifically regulated in erythroid cells, contain GATA1 binding sites [Podkolodnaya and Stepanenko, 1997]. In the ESRG-TRRD database, a description is contained of more than 100 experimentally detected binding sites of this factor. From each GATA-1 binding site described in the ESRG-TRRD database, there are the links to the software program RGSiteScan (http://wwwmgs.bionet.nsc.ru/Programs/yura/rgscan1.html) [Kolchanov, 1999a], which permits to recognize binding sites of 37 transcription factors, including GATA1 transcription factor binding site, in an arbitrary nucleotide sequence. Besides, descriptions of some transcription factor binding sites (in particular, GATA1, AP-1, YY1, and some others) have the links to one more module of the GeneExpress system, the SAMPLES database [Vorobiev et al., 1998]. From this database, a user may extract the samples of functional sites and other biologically significant sequences. In the SAMPLES database (http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/), the sets of sequences of binding sites of about 50 transcription factors are stored. Among them are the sets of some factors important for erythroid-specific transcription regulation, e.g., GATA1, NF-E2, and EKLF, which are built on the basis of information extracted from the ESRG-TRRD database.

In the GeneNet module [Kolpakov et al., 1998] of the information system GeneExpress, the fragment of the gene network is represented, which describes the processes occurring in erythroid cell differentiating under the action of erythropoietin. At present, this fragment contains 171 components, including 32 genes, 51 proteins, 11 low molecular weight components, 87 relations. Figure 1 demonstrates the scheme generated by the software

program GeneNet Viewer on the base of formalized information contained in the GeneNet database (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/).

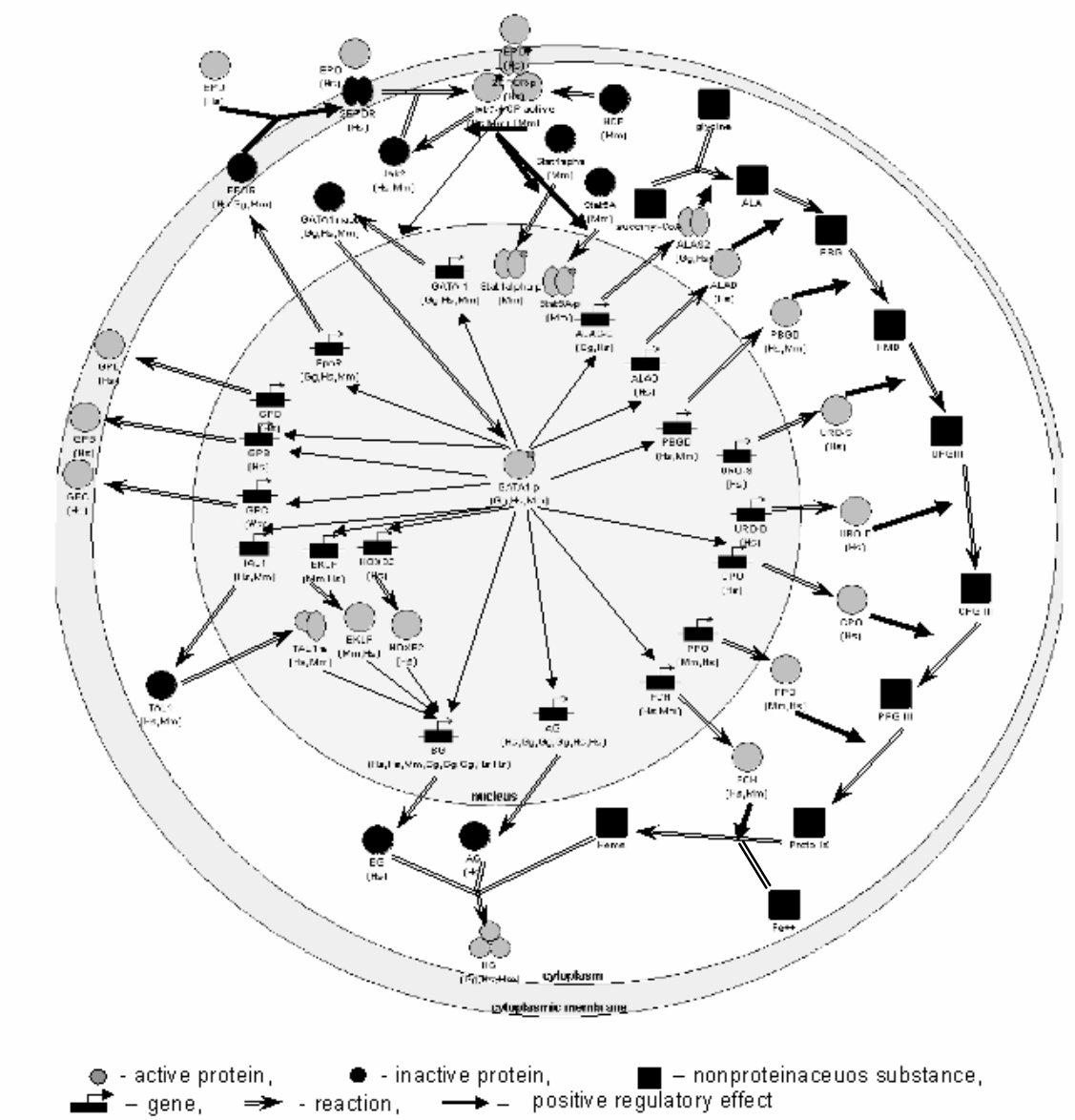


Figure 1. Processes occurring in erythroid cell, which differentiates under the action of erythropoietin (the scheme is generated by the program GeneNet Viewer on the basis of formalized information accumulated in GeneNet database).

The molecule of erythropoietin by means of binding with its own receptor causes its dimerization. After homodimerization of receptor induced by the ligand, an association of the C-terminal cytoplasmic receptor domain with some protein tyrosine kinases (JAK2, Lyn, SHC, SHP-2, etc.), which mediate the transfer of differentiating erythropoietin signal followed by the rapid induction of tyrosine phosphorylation of cell protein substrates together with erythropoietin receptor itself. Next phosphorylation and activation of a seria of transcription factors take place, in particular, STAT family factors, GATA1 factor. Active transcription factors penetrate into the nucleus and activate transcription of respective genes. The presence of GATA-1 binding site in promoter of its own gene produces quick self-enhancement of its transcription byФы the positive feedback mechanism. GATA1 transcription factor through its binding sites in regulatory regions of erythropoietin receptor gene strengthens its transcription. As a consequence, receptor expression increases. By the analogous mechanism, expression is enhanced of such genes as globin genes, surface cell antigens, enzymes of heme biosynthesis pathway. Besides, the same GATA1 transcription factor increases transcription of some transcription factors, which in turn activate transcription of erythroid genes. As a result, expression of cell markers of erythroid differentiation, including glycophorins, heme, and globins, increases.

It should be stressed that cytoplasmic domain of activated erythropoietin receptor may bind to proteintyrosine phosphotase SHP1. As a consequence, dephosphorylation of JAK2 kinase arises, which is followed by decay of signal transduction from activated erythropoietin receptor.

Analysis of data accumulated in the gene network fragment considered demonstrates that stability of differentiation process in erythrocytes depends upon the whole integrity of regulatory mechanisms, which provides high-reliability of erythropoietin signal transduction in a cell and has provision for fine tuning of this process. Let us note the most important components of this regulation.

1. For transduction of positive, stimulating signal of the erythropoietin via its membrane receptor various signal molecules may participate, thus providing amplification of its signal in a cell.
2. Negative feedback acting through the negative regulator – proteintyrosinphosphatase SHP1, suppresses excessive signals from activate erythropoietin receptor.
3. Erythropoietin, by binding to its receptor, switches on the signal transduction pathway, which leads to phosphorylation, acetylation, and activation of some transcription factors, foremost GATA1. Activated GATA1 enhances transcription of erythroid genes including the genes encoding enzymes of heme biosynthesis, globins biosynthesis, erythropoietin receptor and the GATA1 factor itself.
4. There exist two contours of positive feedbacks enhancing erythropoietin signal at the level of transcription: (i) autoregulation of GATA1 transcription factor expression; (ii) activation of erythropoietin receptor gene transcription by the GATA1 factor, followed by enhancement of expression of receptor, transferring the erythropoietin signal to the GATA1 transcription factor gene.
5. Activated GATA1 transcription factor increases erythropoietin signal transduction, by means of some transcription factors activation, these transcription factors together with the GATA1 factor acting in regulation of erythroid-specific gene transcription.

In conclusion, organization of the gene network considered is completely corresponds to the main principles of gene network construction [Kolchanov et al., 2000], that is, (i) there exists a vast variety of molecular mechanisms providing the functioning of feedbacks; (ii) the presence of a "central" gene supporting coordination of functions of the rest components of the network; (iii) cassette mechanism of activation of a large group of genes by a single transcription factor; (iv) the presence of regulatory contours with positive and negative feedbacks that provide autoregulation of the network.

Further development of this informational resource includes increase in number of gene descriptions in the ESRG-TRRD section; more complete description of the pathways transferring the signal from erythropoietin receptor, along with description of regulatory mechanism of heme action; and development of a gene network on iron metabolism in an erythroid cell.

Acknowledgements

This work is partially supported by Russian State Program "Human Genome", Russian State Committee on Science and Technology, Russian Foundation for Basic Research (grants Nos 98-04-49479, 98-07-91078, 99-07-90203, 00-04-49229, 00-07-90337). The authors are grateful to G.V. Orlova for translation of the manuscript into English.

References

1. Kolchanov, N.A., Ananko, E.A., Kolpakov, F.A., Podkolodnaya, O.A., Ignatieva, E.V., Goryachkovsky, T.N., Stepanenko, I.L. (2000) Gene networks. // Molekularnaya biologia, in press (in Russian).
2. Podkolodnaya,O.A., Stepanenko,I.L.(1997) Mechanisms of transcription regulation of erythroid specific genes. // Molekularnaya biologia, T 31, 4, 617-683 (in Russian).
3. Kolchanov,N.A., Ponomarenko,M.P., Frolov,A.S., Ananko,E.A., Kolpakov,F.A., Ignatieva,E.V., Podkolodnaya,O.A. et al., Integrated databases and computer systems for studying eukaryotic gene expression. (1999a). // Bioinformatics. 15, 669-686.
4. Kolchanov,N.A., Ananko,E.A., Podkolodnaya,O,A,, Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., et al., (1999b) Transcription Regulatory Regions Database (TRRD):its status in 1999. // Nucleic Acids Res. 27, 303-306.
5. Kolpakov,F.A., Ananko,E.A., Kolesov,G.B., Kolchanov,N.A. (1998) GeneNet: a database for gene networks and its automated visualization. // Bioinformatics, 14, 529-537.
6. Vorobiev,D.G., Ponomarenko,J.V., Podkolodnaya,O.A.(1998) A samples and aligned: databases for functional site sequences. // Proc. I Intern. Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Russia, 1998, p. 58-61.

SELEX_DB: AN ACTIVATED DATABASE ON DNA/RNA SEQUENCES OBTAINED IN SELEX-EXPERIMENTS

Ponomarenko J.V., Orlova G.V., Ponomarenko M.P., Lavryushev S.V., Zybova S.V., Frolov A.S.

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

Keywords: database, site recognition, DNA, RNA, SELEX-protocol, computer analysis

Resume

Motivation:

During the last decade, novel SELEX-technologies have been developed for revealing high affinity of DNA- and RNA-sequences to various proteins, peptides, and other organic molecules. These sequences could be used in genome annotation.

Results:

The novel database, SELEX_DB, has been developed. This database accumulates DNA/RNA sequences of sites extracted by means of SELEX-technologies out of the pool of randomized sequences. In addition, SELEX_DB contains computer software for recognition of functional DNA/RNA sites.

Availability:

http://wwwmgs.bionet.nsc.ru/systems/Selex

Introduction

Functional site recognition is one of the key aspects of genomic DNA annotation. A huge number of methods have been developed so far to address this problem. The most widely used are those using weight matrices (1-8). For storage of such matrices, the specialized databases were developed, including TRANSFAC (5), IMD (6), RegulonDB (8), etc.

However, over the last decade, the novel SELEX-technologies have been designed for identification *in vitro* of high affinity DNA and RNA sequences to different proteins, peptides, and small organic molecules (for review, see 9-11). Among these technologies are the following: SELEX (Systematic Evolution of Ligands by Exponential enrichment) (12, 13), SAAB (Selected And Amplified Binding site imprint assay) (14), REPSA (Restriction Endonuclease Protection Selection and Amplification) (15), CASTing (Cyclical Amplification and Selection of Targets) (16) and other binding site selection procedures.

Selected high affinity sequences are widely used for recognition and prediction of activities (17, 18). Weight matrices constructed on the base of these sequences are also stored in the databases: TRANSFAC (5), IMD (6), etc. These matrices are used by the programs for site recognition, including TESS (4), MatInspector (2), SIGNAL SCAN (7), etc., together with the matrices calculated for the natural variants of sites. However, the samples of natural sites are more heterogeneous than the sequences selected *in vitro*. Moreover, the experimental conditions are also important. For example, HEN1 protein binding sites are characterized by different consensuses obtained by *in vitro* and *in vivo* experiments (20).

Given current advance in sequencing whole genomes, combinatorial methods will be important in the next generation of studies, thus making the bridge between raw sequence data and actual biological processes. At present, enormous starting libraries are used in different SELEX processes and contain up to 10^{14}–10^{15} sequences (10). Naturally, this information needs to be collected into public databases available via the Internet.

To aim this problem, we have developed a novel database SELEX_DB accumulating the DNA/RNA site sequences selected by SELEX-technologies out of the pool of randomized sequences. The database accumulates also the programs for recognition of functional DNA/RNA sites on the basis of weight matrices stored in the database.

The site sequences stored in the database may be used as independent control data for development of novel methods for site recognition. In addition, this database can be used for designing of novel experiments applying SELEX-technologies.

Format of the database

A database entry corresponds to a single experiment. As an example, an entry containing the information on *in vitro* selected YY1 transcription factor binding sites from a pool of 18 bp random sequences (21).

The entry description is based on 27 fields: **AC**, an accession number of an experiment; **ID**, identifier; **DA**, **DT**, dates of creation and of the last update; **FV**, release number; **MN**, name of an entry; **CR**, name of an annotator (lonked to SCIENTIST database); **NF**, name of a ligand; **OS**, organism; **OC**, taxon; **TE**, templates for amplification; **EX**, type of an experiment; **EC**, experimental conditions (*in vitro* or *in vivo*); **RF**, reference to the literature source (linked to SELEX_BIB database); **KW**, keywords; **NS**, sequence quantity; **AA**, aligned sequences as they are represented in the original paper; **WA**, **WT**, **WG**, **WC**, weight matrix; **CN**, consensus; **DR**, links to the other databases; **WW**, links to recognition programs; **NM**, number of sequences in the set; **SQ,** sequence; **CC,** annotator's comments.

Application of SELEX_DB for genome annotation

To activate SELEX_DB information, the supplementary database SELEX_TOOLS has been developed by analogy to technology applied by the authors earlier in the databases MATRIX (21), ACTIVITY (22) and B-DNA-FEATURES (23). For a fixed functional site with the known weight matrix documented in the fields of the SELEX_DB (**WA**, **WT**, **WG**, **WC**), the C-encoded procedures recognizing the site are stored within the database SELEX_TOOLS. 15 recognition programs are being generated. Namely, seven procedures calculate the weight sums, e.g., homology score (24), matrix similarity (2), etc., seven procedures – weighting consensus match scores (i.e., by Mahalanobis distance, by information content (25), etc.). There is an integrated procedure averaging 14 procedures given above. Thus, a user may choose an approximation, which better suits the particular biological task. Each out of 15 procedures is documented by false positive and negative error rates (Fig. 2A, fields **ST** and **NT**, respectively) and by the hystogram of the score calculated over the site sequences versus random sequences (field **FG**). A user may exploit the chosen procedure in two modes: (i) "on-line" mode, by activation the field "**WW RECOGNITION**" or (ii) "off-line" mode, by extracting the C-codes of this program (the field **C-CODE**) in order to incorporate them into the user's software. This is the novelty of our database.

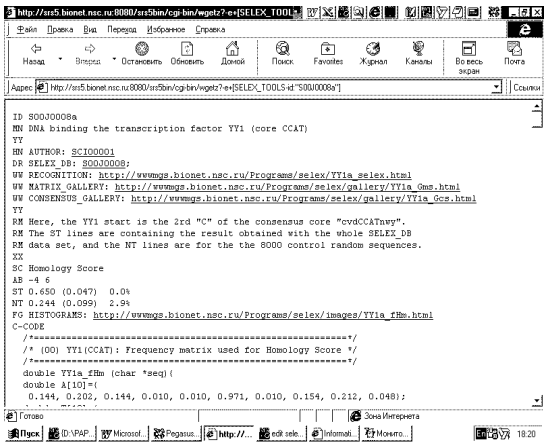
Fig. 2 exemplifies an activation of the SELEX_DB by the example of YY1 transcription factor binding sites (20). By clicking the field "DR, SELEX_TOOLS; S00j008a", one may achieve the entry S00J0008a of the database SELEX_TOOLS (Fig. 2A) and thus apply the C-coded

```
ID S00J0008
XX
AC BS_YY1
XX
DA 13/04/99
DT 13/04/99
FV 1.0
XX
MN Selected YY1 binding sites
XX
CR Ponomarenko JV; SCI00002
XX
NF YY1
OS human
OC EUKARYOTA
XX
TE 5'-AACGGTCCCTGGCTAAAC-18(N)-CAGTGTGTGGACTATTAG-3'
EX PCR-assisted binding site selection
EC in vitro
XX
RF Yant SR et al, 1995; RFSJ0008
XX
KW YY1, globin gene, binding site (Medline,GenBank)
XX
CC CCAT binding core
XX
NS SEQUENCE QUANTITY: 175
XX
AA Aligned sequences from paper
   A2;.....CAGAGACACAGACGCCAT
  A17;.............TACAGCCATTATTCCCCA
  A22;.....CAGACTACAATCTACCAT
  A30;...........TGACCGGCGCCATTGTTA
……
……
  G27;.............TACAGCCATATTACTGCA
  G54;.....TATCNTACGTACCTCCAT
XX
WA 34 28 25 14 20 14  0  0 100  0 15 21  4 18 25 16
WT 26 27 18 16  6 18  0  0  0 100 42 71 81 33 29 39
WG 15 27 39 14 14 67  0  0  0  0 23  5  3 31 22 18
WC 25 18 18 56 60  1 100 100  0  0 20  3 12 18 24 27
XX
CN N N N N V D C C A T N W Y N N N
XX
DR SELEX_TOOLS: S00J0008a
WW RECOGNITION: http://wwwmgs.../selex/YY1a_selex.html
XX
CC ACAT binding core
XX
NS SEQUENCE QUANTITY: 14
XX
AA Aligned sequences from paper
  A12;...........CGGAGACATTTTTGAGTA
  A14;...........GGTAGACATATTCGGGTA
……
……
  F41;...CATCAGGACGGCAGACAT
  G53;....CAGATTAAGGCCGACATT
XX
WA 25 33  8 15 43  0 100  0 100  0 17  0  0 10 11 29
WT 12 17  8 15  0  0  0  0  0 100 67 100 100 30 22 14
WG 38 42 83 23  7 100  0  0  0  0 8  0  0 40 56 29
WC 25  8  0 46 50  0  0 100  0  0 8  0  0 20 11 29
XX
CN N D D N M G A C A T N T T N N N
XX
DR SELEX_TOOLS: S00J0008b
WW RECOGNITION: http://wwwmgs.../selex/YY1b_selex.html
XX
NM A2
SQ CAGAGACACAGACGCCAT
NM A17
SQ TACAGCCATTATTCCCCA
……
NM G53
SQ CAGATTAAGGCCGACATT
//
```

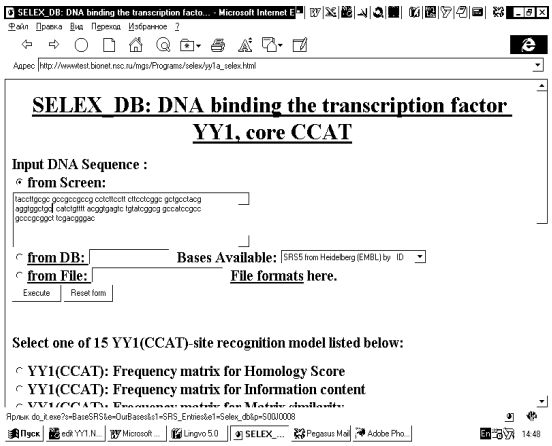
Figure 1. An example of entry in the SELEX_DB database.

program for recognition of transcription factor YY-1 binding sites with the core "CCAT".

A)



B)



C)

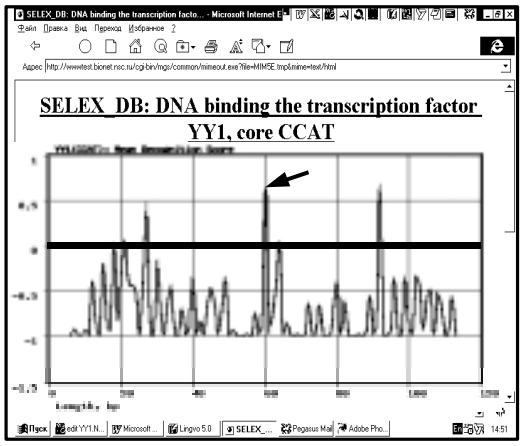


Figure 2. Activation of SELEX_DB: A) C-codes of procedures for recognition of YY1 binding site recognition in SELEX_TOOLS; B) input Web-form for application of recognition procedures to an arbitrary sequence; C) the result of recognition – recognition function value in dependence on the sequence position ("0" denotes the beginning of a sequence).

Activation of the entry "WW RECOGNITION" enables to implement C-procedures for recognition of this site for an arbitrary sequence input by a user (Fig.2B). In Fig. 2C, the resulted YY-1 recognition score profile is shown within the region in-between 7805-7924 positions of Moloney murine leukemia virus gene documented in (EMBL: J02255, REMLM). The peak marked by arrow corresponds to experimentally detected natural YY1 transcription factor binding site in positions 7860-7868, which is documented by the number R01149in the database TRANSFAC (5). Successful recognition of YY1 site in the case considered may be viewed as independent control of the procedure suggested, because under its construction, the natural sites were not considered. Thus, SELEX_DB is directly applicable to analysis of genome sequences.

The other way of SELEX_DB activation is the usage of SRS-formatted keywords SRS (26). By example, by the standard SRS-query with the keyword "DNA-binding", a user may retrieve the entry S00J0008 shown in Fig. 1 and to use the recognition procedures for the YY1 transcription factor binding sites as described above (Fig. 2). Besides, a user may exploit the keyword query generator (27) and to provide an automated search in MEDLINE and GeneBank databases by keywords documented in the field (Fig. 1). For this purpose, it is necessary to click the database name at the end of KW field in SELEX_DB. As a result, the SELEX_DB-related papers will be retrieved, or genomic sequences related to combination of the keywords indicated.

According to the "activated database" approach we have earlier introduced and used (21-23, 28), SELEX_DB has designed to be simultaneously (i) a database, and (ii) the software recognizing the functional DNA/RNA sites, and, also, (iii) a navigator generating and searching for queries through the related databases. This is ,indeed, the novelty of the present work.

Acknowledgements

The work was partially supported by the Russian Foundation for Basic Research (grants Nos. 98-07-910126, 98-07-9107).

References

1. Bucher, P. (1990) *J. Mol. Biol.*, **212**, 563-578.
2. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) *Nucleic Acids Res.*, **23**, 4878-4884.

3. Ghosh, D. (2000) *Nucleic Acids Res.,* **28**, 308-310.
4. Stoeckert Jr, C.J. *et al.* (1999) *Nucleic Acids Res.,* **27**, 200-203.
5. Wingender, E. *et al.* (2000) *Nucleic Acids Res.,* **28**, 316-319.
6. Chen, Q., Hertz, G. and Stormo, G. (1995) *Comput. Applic. Biosci.,* **11**, 563-566.
7. Prestridge, D.S. (1996) *Comput. Appl. Biosci.,* **12**,157-160.
8. Salgado, H. *et al.* (2000) *Nucleic Acids Res.,* **28**, 65-67.
9. Werstuck, G. and Green, M.R. (1998) *Science*, **282**, 296-298.
10. Gold, L. et al. (1997) Proc. Natl. Acad. Sci. USA, **94**, 59-64.
11. Roberts,R.W. and Ja,W.W. (1999) *Curr. Opin. Struct. Biol.,* **9**, 521-529.
12. Tuerk, C. and Gold, L. (1990) *Science,* **249,** 505-510.
13. Ellington, A.D. and Szostak, J.W. (1990) *Nature,* **346**, 818-822.
14. Blackwell, T.K. and Weintraub, H. (1990) *Science*, **250**, 1104-1110
15. Hardenbol, P. *et al.* (1997) *Nucleic Acids Res.,* **25**, 3339-3344.
16. Wright, W.E., Binder, M. and Funk, W. (1991) *Mol. Cell Biol.,* **11,** 4104-4110.
17. Barrick, D. *et al.* (1994) *Nucleic Acids Res.,* **22**, 1287-1295.
18. Liu, H.-X., Zhang, M. and Krainer, A.R. (1998) *Genes Dev.,* **12,** 1998-2012.
19. Yant, S.R. *et al.* (1995) *Nucleic Acids Res.,* **23,** 4353-4362.
20. Brown, L. and Baer, R. (1994) *Mol. Cell Biol.,* **14,** 1245-1255.
21. Ponomarenko, M.P. *et al.* (1999) *Bioinformatics,* **15,** 7/8, 631-643.
22. Ponomarenko, M.P. *et al.* (1999) *Bioinformatics,* **15,** 7/8, 687-703.
23. Ponomarenko, J.V. *et al.* (1999) *Bioinformatics,* **15,** 7/8, 654-668.
24. Mulligan, M.E. *et al.* (1984) *Nucleic Acids Res.,* **12**, 789-800.
25. Schneider, T.D. *et al.* (1986) *J. Mol. Biol.,* **188**, 415-431.
26. Etzold, T. and Argos, P. (1993) *Comput. Applic. Biosci.*, **9,** 49-57.
27. Kolchanov, N.A. *et al.* (2000) *Nucleic Acids Res.*, **28**, 298-301.
28. Ponomarenko J.V., Orlova G.V., et al. (2000) *Nucleic Acids Res.,* **28**, 205-208.

STEROIDOGENESIS-CONTROLLING GENE TRANSCRIPTION REGULATION: REPRESENTATION IN TRRD DATABASE

Busygina T.V., Ignatieva E.V., Osadchuk A.V.

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

Keywords: database, TRRD, transcription regulation, hypothalamic-pituitary-gonadal complex genes, hypothalamic-hypophysial-adrenocortical complex genes, steroidogenic factor 1

Resume

Motivation:

One of the topical goals in modern bioinformatics is to reveal coordinated gene expression mechanisms. The purpose of this study is to clear up the mechanisms of coordinated expression of steroidogenesis-controlling genes.

Results:

ES-TRRD database subsection on transcription regulation of steroidogenesis-controlling genes was developed. A set of transcription factors regulating expression of steroidogenesis-controlling genes that are stored in the ES-TRRD is revealed. Consensus for SF-1 transcription factor binding site was found, this enabling to reveal potential binding sites of this factor within 5'-flanking regions of 2 murine genes encoding microsomal enzymes, 3βHSDI and P45017α. Knowledge on the protein-protein interactions of the SF-1 factor with the other transcription factors and proteins involved in gene transcription regulation was classified.

Availability:

ES-TRRD is available via the Internet by the address http://www.bionet.nsc.ru/trrd/es-trrd/.

Introduction

Steroid hormones regulate vital organism functions. Among these functions are: reproductive function (progestins, estrogens, androgens), anti-stress response (glucocortocoids), and salt balance support (mineralocorticoids). Steroid hormones are synthesized mainly in adrenal cortex, testis, and ovaries. The precursor of all steroid hormones synthesis is cholesterol. The variety of biosynthesis pathways in different tissues is provided under the action of various enzymes localized both at endoplasmic reticulum membranes (3βHSD, P45017α, P450c21, 17βHSD, P450arom) and in mitochondria (P450scc, P45011β, P450aldo). Except enzymes, the proteins with some other functions such as StAR (steroidogenic acute regulatory protein) and ADX (adrenodoxin) are important for steroid hormone biosynthesis. Biosynthesis and secretion of steroid hormones in endocrine glands is controlled by pituitary hormones, i.e., adrenocorticotropin (POMC derivative) and gonadotropic hormones. In turn, pituitary hormone production depends on the level of the so called releasing factors, which are synthesized in the upper

Table 1. Genes controlling steroidgenesis and represented in ES-TRRD.

Genes encoding enzymes and other steroidgenesis factors
StAR (*A00584^{m}, ***A00488^{h}); ADX (A00860^{h}, A00733^{b}); Cyp11A (*A00561^{m}, **A00496^{r}, *A00497^{b}, *A00498^{h}, A00499^{s}); 3βHSDII (*A00858^{h}); 3βHSDII (*A00858^{h}); Cyp17 (A00490^{m}, *A00585^{b}, **A00565^{r}); Cyp19 (*A00591^{r}, *A00152^{h}); Cyp21A1 (***A00790^{m}); Cyp21B (A00588^{h}); Cyp11B1 (A00583^{h}, *A00587^{r}, *A00789^{b}); Cyp11B2 (*A00586^{h})
Genes encoding hormones, their precursors, and releasing factors
GnRH (A00933^{r}, A00934^{m}, A00935^{h}); CRH (A00938^{s}, A00859^{h}); LHB (**A00857^{hrs}, *A00628^{r}); CGB (A00939^{h}); GTHIIB (**A00785^{cs}); GHA (A00791^{m}, *A00056^{h}); POMC (A00430^{r}); OT (*A00630^{b})
Genes encoding hormone receptors
GnRHR (*A00792^{m}); PRLR (*A00629^{r}); ACTHR (***A00929^{h}, *A00786^{m}); LHR (*A00783^{r})
Other genes
LeyIL (***A00784^{m}); MIS (A00781^{m})
Genes encoding transcription factors
SF-1 (A00549^{m}, *A00548^{r}); DAX1 (***A00732^{m})

Notes: in brackets are given accession numbers from the database; denotations of species are as follows: ^{b} - bovine ^{cs} - chinook salmon ^{h} - human ^{hrs} - horse ^{m} - mouse ^{r} - rat ^{s} - sheep; number of asterisks corresponds to the number of SF-1 transcription factor binding sites in gene regulatory regions.

portion of the brain - hypothalamus. Releasing factor receptors and pituitary hormone receptors participate in signal transduction from the cell membrane into cytoplasm.

In order to analyze the mechanisms of genetic control of steroid hormones biosynthesis, we have developed a subsection of ES-TRRD database. This subsection includes the data on transcription regulation of genes encoding enzymes, transport proteins, hormones, hormone precursors, releasing factors, receptors of hormones and releasing factors, along with the genes encoding transcription factors, which regulate all above listed genes (Table 1).

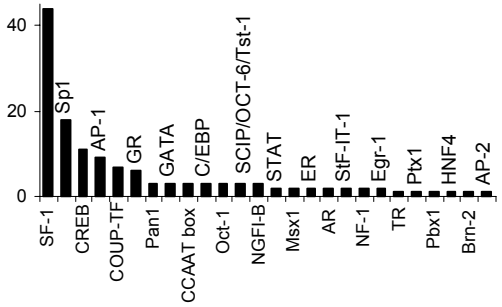


Figure 1. Number of transcription factor binding sites of various types in regulatory regions of steroidgenesis-controlling genes stored in ES-TRRD

Results and discussion

1. Informational content of subsection of the ES-TRRD database on steroidgenesis-controlling genes transcription regulation

We have collected the data on gene transcription regulation in different species: mouse (14), human (13), rat (10), bovine (5), sheep (2), horse (1) (Table 1). More than 65 regulatory regions (i.e., promoters, enhancers, etc.) were characterized along with about 200 transcription factor binding sites. This information is obtained on the base of annotating more than 250 scientific publications.

2. Gene regulatory regions

2.1. Transcription factors regulating expression of steroidgenesis-controlling genes Following the data accumulated in ES-TRRD, more than 40 various transcription factors are involved in steroidgenesis-controlling gene expression regulation. As can be seen from Fig. 1, among the most frequently occurring transcription factor binding sites are the sites of the following types: SF-1, Sp1, CREB, Ap1, COUP-TF, and GR. The factors interacting to these sites perform various functions. Glucocorticoid receptors (GR) mediate the impact of glucocorticoid hormones on the gene expression in correspondance with glucocorticoids concentration in blood, thus realizing the mechanism with the negative feedback. Sp1 is a ubiquitous factor that is involved in regulation of expression of many genes. The factors interacting with CREB and Ap1 sites participate in gene expression regulation in response on alteration of the level of inner-cellular mediators, which, in turn, provide the transduction of signals, which are percepted by cell membrane receptors. COUP-TF factors from the family of nuclear receptors are known as negative transcription regulators. Lastly, SF-1 factor from the family of nuclear receptor genes is known as regulator of development and functioning of hypothalamic-pituitary-gonadal complex and adrenal cortex [Luo,X., 1999].

2.2. SF-1 factors binding sites localization SF-1 transcription factor binding sites are found in regulatory regions of 29 out of 45 genes stored currently in subsection described (Table 1). The genes that are expressed under the control of SF-1 may contain either single or several SF-1 transcription factor binding sites. Examples of SF-1 binding site localization in regulatory regions of five groups of homologous genes are illustrated in Fig. 2. Most sites of this type are located within the region –200/-1 relatively transcription start [Ignatieva, E.V. et al., 2000].

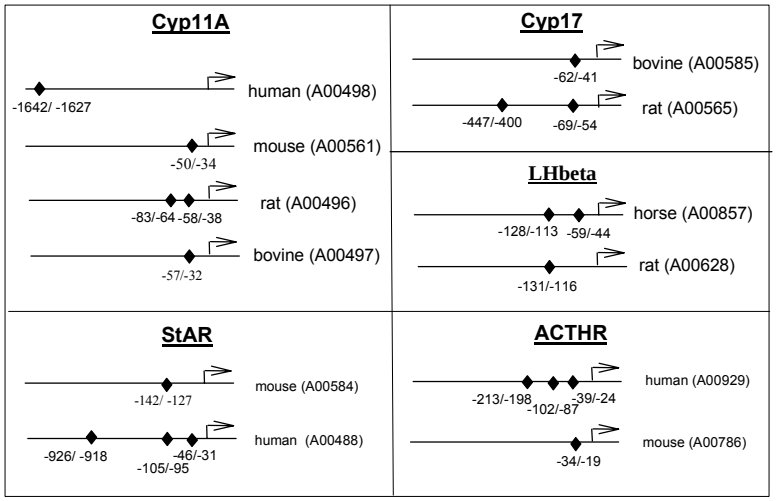


Figure 2. SF-1 transcription factor binding sites localization in regulatory regions of identical genes in different species: cytochrome P450 cholesterol side chain cleavage gene (CYP11A), steroidogenic acute regulatory protein gene (StAR), cytochrome P-450 17alpha hydroxylase/C17-20 lyase gene (CYP17), beta subunit luteinizing hormone gene (Lhbeta), adrenocorticotropin receptor gene (ACTHR);
⇒ marks transcription start.

2.3. SF-1 site consensus Based on information accumulated in ES-TRRD, we have detected a consensus of SF-1 transcription factor binding site with the length of nine nucleotides, TCAAGGTCA (Table 2).

3. Analysis of the mechanisms of genetic control of endocrine function in murine Leidig cells Earlier experiments in mice of 6 inbred strains performed in Laboratory of

endocrine genetics of IC&G SB RAS, revealed strongly coordinated correlative variability in activities of four microsomal steroidgenic enzymes encoded by three genes (3βHSD, Cyp17, and 17βHSD) [Osadchuk,A.V., Svechnikov,K.V., 1998]. Diallele analysis of cAMF and substrate-dependable testosterone production by Leidig cells in mice of above-mentioned strains and their reciprocal F1 hybrids has revealed that the parameters analyzed are under the coordinated and polygene control.

A model supposing existence a four-loci genetic system determining coordinated inherited variability in mouse Leidig cells endocrine activity was developed [Osadchuk,A.V. et al., 1999].

Based on the data accumulated in the ES-TRRD database on wide distribution of SF-1 binding sites in regulatory regions of steroidgenesis-controlling genes, together with the contemporary knowledge on the key role of SF-1 transcription factor in this group of genes transcription regulation, we have supposed that the factor considered may play the role of one out of four loci supporting inherited cooordinated variability in mouse Leidig cells. Until now, neither experimental evidence was obtained that could verify the impact of SF-1 factor in expression regulation of murine genes 3βHSDI, Cyp17, 17βHSD. We have made the prediction of potential SF-1 binding sites in 5'-flanking regions of the genes coding for 3βHSDI and Cyp17. The 3βHSDI gene nucleotide sequence was taken from the literature [Clarke T.R. et al., 1996]. Cyp17 gene sequence is extracted from the EMBL database (S41708). Notably, Cyp17 start transcription site position was calculated on the basis of data stored in TRRD about ARE (androgen responsive element) localization relatively transcription start (identification number of a gene in TRRD is A00490). The nucleotide sequence of the 5' flanking region of the third gene, CYP11A, is still unavailable in mouse.

Table 2. SF-1 transcription factor binding sites in regulatory regions of genes described in TRRD.

Name of a gene	Site number in TRRD	Site sequence	Localization of a site*
MCAD^{h}	S2327	atgTCAAGGCCGtgacccgtgta	+138/+160
Cyp11^{h}	S2674	aggcTCAAGGTCAtca	-1642/-1627
Cyp17^{b}	S3074	aaagtcaaggAGAAGGTCAggg	-62/-41
GHIIB^{chs}	S3696	aaagTAGAGGTCAgga	-175/-160
LHB^{ho}	S3988	gagGCAAGGCCActgg	-59/-44 (R)
Cyp19^{h}	S3313	taCCAAGGTCAgaaat	-135/-120
3βHSDII^{h}	S3992	gagtTCAAGGTAAtaa	-68/-53
DAX1^{m}	S3292	ttTCGAGGTCAtggcca	-131/-115
LeyIL^{m}	S3697	gactTCAAGGTCCcaa	-144/-129
LeyIL^{m}	S3699	cccgCCAAGGCCCatg	-65/-50
MIS^{m}	S3702	cccCCAAGGTCAcctt	-95/-80
Cyp17^{r}	S2649	acgTCAAGGTGAcaat	-69/-54
Cyp11A^{r}	S2682	agggGGGAGGTCAacactcc	-83/-64
Cyp11B1^{r}	S3044	atTCAAGGTTCcacaa	-299/-284
PRLP^{r}	S3063	caggCCAAGGTCAaac	-680/-665
Cyp19^{r}	S3065	ctcCCAAGGTCAtcct	-85/-70
ACTHR^{h}	S4367	ttatTCAAGGTAAtga	-102/-87
ACTHR^{h}	S4368	cggcCCAAGGTCCact	-39/-24
LHR^{r}	S4347	atTCCAGGTCAaggaa	-181/-166
Cyp11A^{m}	S3071	cagcTCAAGGCTAagag	-50/-34 (R)
GHIIB^{chs}	S3694	ttaTCAAGGTCCaagc	-366/-351 (R)
LHB^{b}	S2791	agaggcagACAAGGTCAgggagagg	-133/-109 (R)
Cyp11A^{b}	S2680	tcaccagcTCAAGGCTAagtgagaag	-57/-32 (R)
OT^{b}	S3062	gggTCAAGGTTAtgtc	-166/-151 (R)
LHB^{ho}	S3986	cggACAAGGTCAggga	-128/-113 (R)
StAR^{h}	S2638	gtTCAAGGTCAaa	-928/-916 (R)
StAR^{h}	S2639	ggggTCAAGGATAga	-107/-93 (R)
Cyp11B2^{h}	S3045	CGAAGGTCAaggctggag	-129/-112 (R)
DAX1^{m}	S4305	ttggACAAGGGCgcag	-83/-68 (R)
LeyIL^{m}	S3698	cgagTCACGGTCAggg	-114/-99 (R)
GNRHR^{m}	S3754	ccTGAAGGCCAagtgt	-250/-235 (R)
Cyp11A^{r}	S2683	agcTCAAGGCTAagagaggag	-58/-38 (R)
LHB^{r}	S3050	cagACAAGGTCAgaaa	-131/-116 (R)
ACTHR^{h}	S4366	tagTCAAGGTTActtc	-213/-198 (R)
KOH-CEHCYC		TCAAGGTCA	

(R) – the sequence is located in complementary chain.

Denotations within gene names: ^{h} - human, ^{m} - mouse, ^{r} - rat, ^{b} - bovine, ^{chs} - chinook salmon, ^{h} – horse

* relatively transcription start

There were detected nine potential sites such that they differ from consensus at most for 2 nucleotides. The sequence of the Cyp17 gene contains 6 potential SF-1 transcription factor binding sites (Fig. 3). One of them (-52/-44) is located in direct orientation, whereas the other five sites (-505/-497; -435/-427; -337/-329; -285/-277; -279/-271) are oriented in reverse mode. In-between (-285/-271) positions, we have found two intersecting SF-1 sites. The 5' flanking sequence of the 3βHSDI gene has three potential SF-1 transcription factor binding sites (Fig. 3). One potential site, (-84/-73), is located in direct orientation and the rest two sites, (-117/-109; +86/+94), are reversed. Notably, two out of three potential binding sites detected are located in the 5' flanking region of a gene, whereas the third site is found within the first exon.

The presence of potential SF-1 transcription factor binding sites in regulatory regions of mouse Cyp17 and 3βHSD genes gives evidence that this factor could be considered as one out of four loci determining coordinated inherited variability of hormone activity in mouse Leidig cells.

According to the data stored at present in the ES-TRRD database, along with experimental data, which were not accumulated in the ES-TRRD due to specificity of its format, SF-1 is able to regulate gene expression by interaction with the other transcription factors and proteins. The main types of interactions, proteins interacting with SF-1 factor and genes regulated in such a manner are shown in Table 3. The data summarized in the Table may serve as a foundation for the further theoretical analysis of gene transcription regulation mechanisms determining hormone activity of Leidig cells in mouse.

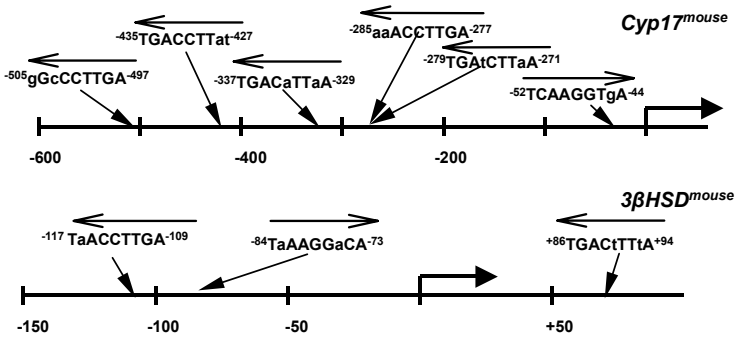


Figure 3. Potential SF-1 transcription factor binding sites in regulatory regions of mouse Cyp17 and 3βHSD genes. Potential sites differing from consensus at most for two nucleotides are shown. Nucleotides coinciding with consensus are given in capital letters. Arrows indicate direction of transcription factor binding site (from 5' to 3'). ⊢> denotes transcription start.

Table 3. Interactions of SF-1 factor with the other proteins involved in transcription regulation.

Type of interaction	Proteins interacting with SF-1	Names of genes, species-specificity, and number in TRRD / Reference
Composite	SOX9	MIS^{human}(A00988)
Element of	ER	GTHIIB^{salmon}(A00785)
synergetic	GATA-4	MIS^{mouse} (A00781)
type*	EGR-1	LHB^{horse} (A00857), LHB^{rat} (A00764)
	Ptx-1	LHB^{bovine} (A00480)
	CREB/ATF	Cyp11A1^{human} (A00498)
Composite element of antagonistic type *	COUP-TF	DAX1^{mouse} (A00732), CYP19^{human} (A00152), CYP11B2^{human} (A00586), Cyp17^{bovine} (A00585), Cyp17^{rat} (A00565), OT^{bovine} (A00630)
Other protein-protein interactions	cJun	Li,L.A. et al., 1999
	SRC-1	Crawford,P.A. et al., 1997
	N-CoR	Crawford,P.A. et al., 1998
	WT-1	Nachtigal,M.W. et al., 1998
	CBP/p300	Monte,D. et al., 1998
	DAX-1	Nachtigal,M.W. et al., 1998; Crawford,P.A. et al., 1998; Zazopoulos E. et al., 1997
	TFIIB	Li,L.A. et al., 1999

* Detailed description of composite elements of synergetic and antagonistic types is given elsewhere [Kel,O.V. et al., 1995]

Acknowledgments

The authors are grateful to I.V. Lokhova for the technical assistance in development of ES-TRRD section, to V. Levitsky and D. Vorobiev for the help in nucleotide sequence analysis; to G.V. Orlova for translation of the paper into English. The work was supported by the Russian Foundation for Basic Research (grants Nos. - 98-04-49452, 98-04-49479, 98-07-91078, 99-07-90203, 00-04-49229, 00-04-49255, 00-07-90337) and by DOE USA (grant DE-FG02-00ER62893/535228).

References

1. Clarke,T.R., Bain,P.A., Burmeister,M., Payne,A.H. (1996) Isolaton and characterization of several members of the murine Hsd3b gene family. DNA and Cell Biology, V. 15, № 5, 387-399
2. Ignatieva,E.V., Busygina,T.V., Ananko,E.A., Podkolodnaya,O.A., Merkulova,T.I., Suslov,V.V., Pozdnyakov M. (2000) Databases on endocrine system gene expression regulation: informational content and computer analysis. This issue.
3. Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E., Kolchanov,N.A. A compilation of composite regulatory elements affecting gene transcription in vertebrates. (1995) Nucleic Acids Res. 23, 20, 4097-103.
4. Luo,X., Ikeda,Y., Lala,D., Rice,D., Wong,M., Parker,K.L. (1999) Steroidogenic factor 1 (SF-1) is essential for endocrine development and function. J.Steroid.Biochem.Mol.Biol., 69, 13-18.
5. Osadchuk,A.V., Svechnikov,K.V., Ahmerova,L.G. (1999) A four-locus least squares linear model of testosterone production by Leydig cell in mice. Thirteen International Mouse Genome Conference, October 31 - November 3, 1999, Philadelphia, PA, USA Abstract E24
6. Osadchuk,A.V., Svechnikov,K.V. (1998) Genetic control of microsomal enzymes activity in steroidgenesis in Leidig cells of inbred mouse strains. Genetika (Moscow), 34, 9, 1277-1285 (in Russian).

DATABASE ON COMPOSITE REGULATORY ELEMENTS IN EUKARYOTIC GENES (COMPEL)

[1]Kel-Margoulis O.V., [1]Romaschenko A.G., [1]Deineko I.V., [1]Kolchanov N.A., [2]Wingender E., [1]Kel A.E.

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: okel@bionet.nsc.ru
[2]Gesellschaft für Biotechnologische Forschung mbH, Braunschweig, Germany
e-mail: ewi@gbf.de
*Corresponding author

Keywords: transcriptional regulation, composite regulatory elements, relational database, gene-specific regulation, transcription factor, factor-factor interactions, factor-DNA interactions, DNA-binding domains

Resume

Motivation:

Composite regulatory elements contain two closely situated binding sites for distinct transcription factors, and actually are minimal functional units on DNA providing cross-coupling of signal transduction pathways. Both specific factor-DNA and factor-factor interactions contribute to the function of composite elements. Information about structure of known composite elements and specific regulation provided by them appears to be extremely useful for promoter prediction and for applied gene engineering as well. Therefore collection and classification of rapidly accumulating experimental data in a specialized database are needed.

Results:

The format of data presentation and a relational model of the COMPEL database have been developed. Composite elements are classified according to the specific function they provide. WWW search and browse routines were developed for COMPEL release 3.0. Based on the COMPEL collection software has been developed for searching potential composite elements in gene regulatory regions.

Availability:

The COMPEL database equipped with the search and browse tools is available at http://compel.bionet.nsc.ru. The program for searching potential composite elements is available at http://compel.bionet.nsc.ru/FunSite/CompelPatternSearch.html

Introduction

Presently, there are a number of databases on transcriptional regulation available. The COMPEL database emphasizes the key role of specific interactions between transcription factors binding to their target sites providing specific features of gene regulation in a particular cellular content.

Based on the known examples we define a composite element as a minimal functional unit where both protein-DNA and protein-protein interactions contribute to a highly specific pattern of gene transcriptional regulation (Kel O.V. et al., 1995; Kel O.V. et al., 1997; Kel-Margoulis et al., 2000). Thus interacting factors may differ by the structure of DNA-binding, activation, oligomerization and other domains. Along with structural differences, functional properties of the transcription factors and hence their specific contribution to the transcription regulation may significantly vary. Co-operative action of the transcription factors binding to their target sites within the composite elements results in a new highly specific pattern of gene transcription that can not be provided by factors separately. Composite elements are structural-functional units that provide cross-coupling of gene regulatory pathways, and in particular, cross-coupling of signal transduction pathways (Kel-Margoulis et al., 2000).

There are two main types of composite elements: synergistic and antagonistic ones. In synergistic CEs, simultaneous interactions of two factors with closely situated target sites results in a non-additive high level of a transcriptional activation. Within an antagonistic CE two factors interfere with each other. A number of molecular mechanisms have been suggested for functioning of both synergistic and antagonistic CEs (Kel O.V. et al., 1997).

On the base of information collected in the COMPEL database we have classified composite elements according to their specific contribution to the pattern of gene transcription. The most numerous classes contain CEs providing cross-coupling of different signal transduction pathways.

COMPEL has been developed in a joint effort of the Institute of Cytology and Genetics (Novosibirsk, Russia) and Gesellschaft für Biotechnologische Forschung mbH (Braunschweig, Germany) since 1994. The structure of the previous COMPEL releases has been described earlier (Kel O.V. et al., 1995; Kel O.V. et al., 1997; Wingender et al., 1997; Heinemeyer et al., 1998; Kel-Margoulis et al., 1998; Heinemeyer et al., 1999; Kel-Margoulis et al., 2000). During last two years, the structure of the database has been improved considerably. One important new feature is the link to the EMBL databank. COMPEL is publicly available for non-commercial users and is distributed in three interlinked ASCII flat files. We have also developed search and browse tools that are available via WWW (http://compel.bionet.nsc.ru/).

Classification of the composite elements

We have classified CEs according to the specific transcriptional regulation they provide due to co-operative action of transcriptional factors binding to their target sites. 137 CEs have been classified including 120 CEs of synergistic type and 17 of antagonistic (Kel O.V. et al., 1997; Kel-Margoulis et al., 1998; Kel-Margoulis et al., 2000) (Table 1). The majority of CEs contain at least one binding site for an inducible factor (102 CEs), and a number of CEs contain at least one binding site for a tissue-enriched factor (51 CEs). CEs are classified into five main groups, (examples are shown in Table 1): 1) 64 CEs formed by binding sites for two inducible factors, they provide cross-coupling of signal transduction pathways; 2) 22 CEs formed by binding sites for a tissue-enriched and an inducible factor, they provide tissue-specific responses to inducing signals; 3) 18 CEs formed by binding sites for a tissue-enriched and a constitutive ubiquitous factor, they provide some additional features of the tissue-specific transcriptional regulation; 4) 14 CEs formed by binding sites for an inducible and a constitutive ubiquitous factor, they provide some additional features of the inducible regulation; 5) 11 CEs formed by binding sites for two tissue-enriched factors, they provide some particular tissue-specific regulation.

Table 1.Classification of the CEs according to the specific function they provide.

Functional properties TF1	Functional properties TF2	Gene	Location of CE	COMPEL acc [1)]
1) CEs providing cross-coupling of signal transduction pathways				
c-Ets, Ras-dependent ind.	AP-1, ind. PKC	SR; macrophage scavenger receptor gene, Hs	-65 … -52	C00079
C/EBPβ, ind. IL-6	NF-κB, ind. IL-1 and TNFα	Serum amiloid A2, Hs	-179 … -82	C00100
NF-AT, ind. Ca^{2+}	AP-1, ind. PKC	Interleukin-2, Hs	-287 … -266	C00109
C/EBPβ, ind. IL-6	AP-1, ind. PKC	TNFα, Hs	-107 … -74	C00178
IRF-1, ind. by interferon α и γ	NF-κB, ind. IL-1 and TNFα	Interferon β, Hs	-77 … -55	C00061
2) CEs providing tissue-restricted response to an induction				
HNF-1, hepatocyt.	C/EBPβ, ind. IL-6	β-fibrinogen gene, Hs	-133 … -77	C00095
HNF-3, hepatocyt.	GR, ind. Glucocort.	Tyrosine amino transferase, Rn	-2509... -2430	C00128
HNF-4, hepatocyt.	CREB, ind. CAMP		-3650… -3586	C00129
AML1, T- and myeloid cells	c-Ets, Ras-dependent ind.	T-cell receptor β, Hs	3' enchancer	C00020
Pit-1, pituitary	c-Ets, Ras-dependent ind.	Prolactin, Rn,	-162 … -147 -217… -190	C00137 C00131
3) CEs providing tissue-restricted regulation for that ubiquitous factor is essential				
Myogenin, muscle cells	Sp1, ubiquitous	Acetylcholine receptor α-subunit, Hs	-89 … -47	C00027
Pit-1, pituitary	Sp1, ubiquitous	Growth hormone gene, Hs	-139 … -105	C00038
HNF-1, hepatocyt.	Oct-1, ubiquitous	Large surface antigen, HBV	-86 … -51	C00048
C/EBP α, hepatocytes	NF-Y, ubiquitous	Serum albumin, Mm	-110 … --80	C00069
4) CEs providing inducible regulation for that constitutive factor is essential				
NF-κB, ind. IL-1 and TNFα	Sp1, constitutive	HIV-1 LTR	-90 … -68	C00055
c-Ets, ind. by Ras	Sp1, constitutive		-140 … -127	C00007
C/EBPβ, ind. IL-6	Sp1, constitutive	CYP2D5, Rn	-105 … -83	C00070
Stat 3, ind. IL-6	Sp1, constitutive	Transcription factor C/EBPδ, Mm	-120 … -102	C00179
GR, ind. glucocort.	Oct-1, constitutive	MMTV LTR	-89 … -49	C00043
5) CEs providing some additional features of tissue-restricted regulation				
HNF-4 - liver, gut, kidney	C/EBPα - liver, gut, adipocytes, brain, myelocytes	Apolipoprotein B, Hs	-81 … -52	C00122
C/EBPα - liver, gut, adipocytes, brain, myelocytes	AML1, T- and myeloid cells	M-CSF receptor (c-fms), Hs	-84 … -67	C00145
MEF2A, muscle cells	Myogenin, muscle cells	Muscle-restricted transcription factor MRF4, Rn	-26 … +27	C00120

1) By this acc.number a detailed information about composite element can be found in the COMPEL database (http://compel.bionet.nsc.ru/compel/search.html)
Hs - *Homo sapiens*; Mm - *Mus musculus*; Rn - *Rattus norvegicus*; HBV – human hepatitis B virus; HIV – human immunodeficiency virus type 1; MMTV – mouse mammary tumor virus.

Database structure and WWW interface

The relational model of COMPEL was described earlier (Kel-Margoulis et al., 1998; Heinemeyer et al., 1999). COMPEL has been made publicly available and distributed in three ASCII flat files: **compelement.dat**, **interaction.dat** and **reference.dat** (Kel-Margoulis et al., 2000). A detailed description of the fields is given in the database documentation (*http://compel.bionet.nsc.ru/compel/description.html*). COMPEL is closely linked to other databases on transcriptional regulation, TRANSFAC and TRRD. TRRD contains the field "CE" which refers to COMPEL through composite element accession number (Kolchanov et al., 1999). The file **interaction.dat** is connected with TRANSFAC FACTOR table. Most of the composite elements are linked to the EMBL databank, and all references are linked to MEDLINE.

WWW search and browse options are available now for COMPEL release 3.0. The browsing is provided through the type of DNA binding domains of transcription factors involved. For example, by clicking to the REL tag one will get a list of all composite elements in COMPEL comprising at least one binding site for a transcription factor containing REL-homology domain (NF-κB or NFATp/c). From this list one could retrieve any individual COMPEL entry supplied with all necessary hyperlinks to interaction entries, references, as well as to the foreign databases: TRRD, TRANSFAC, EMBL and MEDLINE. The search engine enables users to retrieve COMPEL entries by gene name, species, name of transcription factor, DNA-binding domain as well as to make a full-text query. Simple Boolean operations on two terms are available.

Connected programs

Currently, a program for searching potential composite elements in DNA sequences is available on the COMPEL web site (http://compel.bionet.nsc.ru /FunSite/CompelPatternSearch.html). A sequence under study is scanned by this program using all composite elements collected in the COMPEL as individual searching patterns. Several parameters are available restricting the search, such as: maximal mismatches in the cores of site1 and site2 comprising the composite elements, maximal variation of the distance between two sites (by percents), composite score cut-off value. The composite score reflects how well the match coincides with the known example of the composite element in COMPEL. Score function takes into account the number of mismatches in both sites and distance between them. All found matches are directly linked to the COMPEL entries containing the corresponding composite elements.

We have applied the program CompelPatternSearch for 5'regions of genes expressed in activated immune cells (***T***-genes). For this analysis we have chosen a set of CEs that are situated in the regulatory regions of genes expressed in activated T-, B- and myeloid cells and collected in the COMPEL database. The set included the following types of CEs: AP-1/NF-κB, AP-1/Oct, AP-1/NFAT, AP-1/Ets, NF-κB/HMG, NF-κB/IRF, C/EBP-α / AML, C/EBP-α / PU.1, Ets/AML. The frequency of the potential CEs in the 5' regions of ***T***-genes is 3 times higher than in the random sequences with the same nucleotide composition (Table 2). Some interesting examples of the potential CEs found by the program will be presented.

Table 2. Frequency of potential CEs that provide various aspects of lymphoid- and myeloid-restricted transcriptional regulation within different sequences.

Sequences under study	Frequency of potential CEs on 1000 bp
T-genes	1.636
5'regions of T-genes	2.894
Random [A]=[T]=[C]=[G]=0.25.	0.832
Random (nucleotide composition as in 5'regions of T-genes) [A]=0.2865;[T]=0.2615; [G]=0.2196; [C]=0.2323	0.938

Acknowledgements

Different parts of this work were funded by the German Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (FANGREB project and project no. X224.6), by the Russian Ministry of Sciences and the Siberian Branch of Russian Academy of Sciences, by the North Atlantic Treaty Organisation (grant no. 951149), by Volkswagen-Stiftung (I/75941) as well as by BIOBASE Ltd (Braunschweig, Germany).

References

1. Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L., and Kolchanov,N.A. (1998) *Nucleic Acids Res.*, **26**, 362-367.
2. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) *Nucleic Acids Res.*, **27**, 318-322.
3. Kel,A., Kel-Margoulis,O., Babenko,V., and Wingender,E. (1999) *J. Mol. Biol.*, **288**, 353-376.

4. Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E., and Kolchanov,N.A. (1995) *Nucleic Acids Res.*, **23**, 4097-4103.
5. Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E., and Kolchanov,N.A. (1997) *Mol. Biol. (Mosk).*, **31**, 498-512.
6. Kel-Margoulis,O.V., Kel,A.E., Frisch,M., Romaschenko,A.G., Kolchanov,N.A., and Wingender,E. (1998) *Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure*, (BGRS'98), ICG, Novosibirsk, Vol.**1,** 54-57.
7. Kel-Margoulis,O.V., Romaschenko,A.G., Kolchanov,N.A., Wingender,E., and Kel,A.E. (2000) *Nucleic Acids Res.*, **28**, 311-315.
8. Kolchanov,N.A., Ananko,E.A., Podkolodnaya,O.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis, O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busigina,T.N., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N., Romashchenko,A.G. (1999) *Nucleic Acids Res.*, **27**, 303-306.
9. Wingender,E., Kel,A.E., Kel,O.V., Karas,H., Heinemeyer,T., Dietze,P., Knüppel,R., Romaschenko,A.G., and Kolchanov,N.A. (1997) *Nucleic Acids Res.*, **25**, 265-268.

PATHO DB – A DATABASE BRIDGING THE GAP BETWEEN THE DESCRIPTION OF GENE REGULATORY DEFECTS AND CLINICAL APPLICATION

*[1]Pruess M., [2,3]Meinhardt T., [1,2]Wingender E.

[1]Biobase Biological Databases GmbH, Braunschweig, Germany
e-mail: mpr@biobase.de
[2]GBF-Gesellschaft für Biotechnologische Forschung mbH, Braunschweig, Germany
e-mail: ewi@gbf.de
[3]Im Weiher 1-3, Heidelberg,Germany
e-mail: thorsten.meinhardt@novasoft.de
* Corresponding author

Keywords: transcription factors, transcription factor binding sites, database development, regulatory defects, developmental disturbances, oncological diseases, molecular diagnostics

Resume

Motivation:

Databases on regulatory elements are already existing, concerning genes, proteins and/or molecular interactions (for an overview, see Heinemeyer et al., 1999). Often these elements can occur in mutated forms which cause disturbances of the transcriptional control and, thus, result in specific diseases.

To achieve a more detailed insight into genotype-phenotype correlations and thereby gaining a deeper understanding of regulatory mechanisms, we established a database, PathoDB, in which we collect different types of pathologically relevant data and model the appropriate relations.

Results:

The relational database PathoDB has been developed and implemented. It now represents a collection of more than 10400 mutated transcription factors and binding sites and combines information about the underlying molecular defects, protein features and functions and the clinical outcome of gene regulatory defects.

Availability:

PathoDB will be available presumably by the end of 2000. It will be integrated with the TRANSFAC database on transcription factors and transcription factor binding sites, so that information about wildtypes and mutated forms will be accessible through one interface.

Introduction

For the regulation of gene expression, elements like transcription factors and transcription factor binding sites are essential. They operate at the level of transcriptional activation, which is the main level of expression regulation. Especially during development of a multicellular organism, the exact coordination of gene expression is of great relevance. In case of disturbances of this coordination, either by defects of transcription factors or of the binding site, serious dysplasias or other diseases can result.

In cancer development, mutated transcription factors play a key role, too. Some factors normally act as inhibitors of transcription, whereas mutated forms lead to uncontrolled gene expression and thus acquire oncogenic activity. Since oncological diseases are of particular interest in medical and pharmaceutical research, efforts to identify the underlying causes and working mechanisms are necessary. Therefore, information about molecular incidences are collected in PathoDB. In connection with the more clinical information, this can provide a useful resource for gene therapy research purposes. Other than in metabolic diseases, for example, where permanent gene expression is essential for the treatment, permanent expression is not necessary for the treatment of cancer since short term gene expression is sufficient for antitumoral toxicity.

Methods and algorithms

The database has been established under a relational database management system (MS Access). It will be made available as an ASCII flat file version.

Implementation and results

PathoDB has been implemented as a relational database system; it consists of a total of 34 tables, and 12 TRANSFAC tables are directly linked. With this close internal link to the TRANSFAC system, optimal analytic

possibilities are opened up, like links to sites and (interacting) wild type factors, regulating pathways, and cross-comparison of mutated versus wild type entries. To access data beyond PathoDB's primary field of interest, the external databases EMBL (Baker et al., 2000), GeneCards (Rebhan et al., 1997), HGMD (Krawczak and Cooper, 1997), MGI (Blake et al., 1999), OMIM (Online Mendelian Inheritance in Man 2000), and SwissProt (Bairoch and Apweiler, 2000) are connected as well.

Currently, 10450 mutated transcription factors and 19 mutated binding sites are recorded in the database, mainly of human, but some also of murine origin. The factors described range from those that are responsible for early embryonic development such as the Pax factors controling eye, ear or kidney development, to those which are important for tumor suppression such as p53. For each, the given information encompasses the mutated sequence, features, functional properties, the underlying genotype, and the corresponding phenotype. Also, detailed information about methods for molecular diagnostics of the defects is given. The data in PathoDB are mostly selected from original articles and reviews published in peer-reviewed journals. Some data are automatically extracted from a publicly available online-database, the IARC Database of p53 (Hainaut et al., 1998).

Discussion

In PathoDB, we present a new combination of molecular and pathological information concerning mutated regulatory elements in order to make regulatory defects more transparent. Since for an effective implementation of gene therapy a comprehensive knowledge of gene expression is absolutely essential, we try to contribute to this knowledge by a structured representation of the relevant pathological data.

Among other things, a large collection of mutated sequences is given which can be useful for other bioinformatic tools which compare and/or estimate sequences or binding sites. For simulation purposes, the question if a mutated factor could bind to a specific site, for example, and what could happen then, is a very interesting one. Moreover, we plan to provide automatically derived primer sequences to facilitate the molecular diagnosis of diseases.

PathoDB will be part of the TRANSFAC system (Wingender et al., 2000).

Acknowledgements

This work has been supported by a grant of the German Ministry of Education, Science, Research and Technology (BMBF; Project No. 0311640).

References

1. Bairoch,A., Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res., **28,** 45-48
2. Baker,W., van den Broek,A., Camon,E., Hingamp,P., Sterk,P., Stoesser,G., Tuli,M.A. (2000) The EMBL Nucleotide Sequence Database. Nucleic Acids Res., **28,** 19-23
3. Blake,J.A., Eppig,J.T., Richardson,J.E., Davisson,M.T. (2000) The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. Nucleic Acids Res., **28**, 108-111.
4. Hainaut,P., Hernandez,T., Robinson,A., Rodriguez-Tome,P., Flores,T., Hollstein,M., Harris,C.C., Montesano,R. (1998) IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. Nucleic Acids Res., **26**, 205-213.
5. Heinemeyer,T., Chen,X., Karas,H., Kel,A.E., Kel,O.V., Liebich,I., Meinhardt,T., Reuter,I., Schacherer,F. and Wingender,E. (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. Nucleic Acids Res., **27**, 318-322.
6. Krawczak,M., Cooper,D.N. (1997) The Human Gene Mutation Database. Trends Genet., **13**, 121-122.
7. Online Mendelian Inheritance in Man, OMIM (TM) (2000). McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). World Wide Web URL: http://www.ncbi.nlm.nih.gov/omim/
8. Rebhan,M., Chalifa-Caspi,V., Prilusky,J., Lancet,D. (1997) GeneCards: encyclopedia for genes, proteins and diseases. Weizmann Institute of Science, Bioinformatics Unit and Genome Center (Rehovot, Israel). World Wide Web URL: http://bioinfo.weizmann.ac.il/cards
9. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I., Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res., **28**, 316-319.

THE TRANSPATH SIGNAL TRANSDUCTION DATABASE: A KNOWLEDGE BASE ON SIGNAL TRANSDUCTION NETWORKS

*[1]Schacherer F., [2]Choi C., [2]Götze U., [2]Krull M., [1,2]Wingender E.

[1]GBF German Research Center for Biotechnology Braunschweig, Germany
e-mail: frs@gbf.de
[2]Biobase Biological Databases GmbH, Braunschweig, Germany
e-mail: ewi@biobase.de
*Corresponding author

Keywords: signal transduction, signaling networks, database development, object-oriented databases, regulatory defects, oncological diseases

Resume

Motivation:

Knowledge about the principle mechanisms of signal transduction and regulation mechanisms of individual macromolecules in signaling pathways has multiplied in the last decade. It is now growing at a rate that makes it difficult to keep up with (Krauss, 1997). Signal transduction pathways regulate the activity of many transcription factors (Montminy, 1997) and practically all oncogenes encode aberrantly functioning members of such pathways coupled to growth-regulating signals (Egan and Weinberg, 1993).

Results:

TRANSPATH is an information system on gene-regulatory pathways, and an extension module to the TRANSFAC database (Wingender et al., 2000). It focuses on pathways involved in the regulation of transcription factors in different species, mainly human, mouse and rat. Elements of the relevant signal transduction pathways like hormones, receptors, enzymes and transcription factors are stored together with information about their interaction and references in an object-oriented database. All information is validated with references to the original publications. Also, references to other databases are provided (TRANSFAC, Swissprot, EMBL, PubMed and others).

Availability:

The database is available over WWW (http://transpath.gbf.de). There are also graphic maps of selected pathways which allow to access the data by clicking on them. As an exchange format for the data, XML (eXtensible Markup Language) flatfiles can be created from and read into the database. This site also provides an interface to the CSNDB database on signal transduction (Takai-Igarashi and Kaminuma, 1998).

Introduction

Cells, especially those of a complex multicellular organism, have to act and react to each other and to external influences in a well concerted manner. Thus, if we want to understand cellular behaviour and its responses to external signals, or want to influence it in a predictable manner, we have to understand the pathways through which these signals are mediated into and within the cell. In most cases changes in cell behaviour involve the execution of transcriptional events, which are specific for each signal in its cellular context (Hill and Treisman, 1995). Biological signaling pathways also interact with each other to form complex networks. These networks show emergent properties like signal integration accross multiple time scales or self-sustaining feedback loops which are not present in the isolated pathways (Bhalla and Iyengar, 1999).

The huge and ever more rapidly growing amount of signal transduction data demands for a database that stores and organizes this knowledge, providing simple and fast access to the information. The complexity created by the cross-talk between pathways makes it virtually impossible to infer by hand all the consequences that follow after one modifies one part of the network. To this end, computer-aided simulation will have to be used. It can only be successful on the basis of a comprehensive and detailed dataset.

Methods and algorithms

The database has been established under an object-oriented database management system (POET Software 1999). As an interface to the database, Java and Object Query Language (ODMG 1997) are used in servlets, providing access over the WWW. The data is stored as a bipartite hypergraph. To visualize the data, an expanded depth-first graph traversal algorithm (Jungnickel, 1994) is used, which allows for searches that make use of protein family information.

Implementation and results

The core development of the Transpath database is complete. The database system consists of 107 classes, 7 of which are persistently used for data storage. There are about 16000 Molecules and 800 interactions in the database currently and it is updated daily. The molecules were largely imported from the public SWISS-PROT database (Bairoch and Apweiler, 2000) whereas all of the interactions are retrieved from original literature.

Interactions are modeled as reactions with reactants and products and a single enzyme or inhibitor. To enable the system to be used as the basis for simulation it is necessary to include rate constants in the reactions and different entries for different states of a molecule. Alternatively to this mechanistic view, interactions can be stored as activation and inhibition pointers providing a semantic view (similar to that provided by CSNDB) which corresponds to the schematic drawings familiar from the literature.

Queries to the database are conducted via the Internet by submitting names of transcription factors or other signal molecules. The user can choose to view either the encyclopaedic information for the requested molecule or the reaction cascades starting from the molecule.

Discussion

Transpath provides a knowledge base which goes beyond the approach of traditional gene or sequence databases by focusing on the interactions between the stored data items. By building up the signaling network from single interactions instead of using predefined pathways, it becomes possible to explore the pathways through the graph in an unbiased way.

The next step will be to analyze the reaction graph and infer general properties of the signal transduction pathways involved. With more and detailed data, it might also become feasible to run simulations to obtain suggestions for the response behaviour of such networks to extracellular signals, which then may be used to drive experimental research.

Acknowledgements

This work has been supported by a grant of the German Ministry of Education, Science, Research and Technology (BMBF; 01 KW 9629/7). The complete data set of CSNDB was generously provided by T. Takai-Igarashi.

References

1. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res., 28, 45-48.
2. Bhalla,U.S. and Iyengar,R. (1999) Emergent Properties of Networks of Biological Signaling Pathways. Science, 283, 381-387.
3. Catell,R.G.G. andBarry,D.K. (1997) The Object Database Standard: ODMG 2.0. Morgan Kaufmann Pub., San Fransisco.
4. Egan,S.E. and Weinberg,R.A. (1993) The pathway to signal achievment. Nature, 365, 781-83.
5. Hill,C.S. and Treisman,R. (1995) Transcriptional regulation by extracellular signals: mechanisms and specificity. Cell, 80, 199-211.
6. Jungnickel,D. (1994) Graphen, Netzwerke und Algoriothmen. BI Wissenschaftsverlag, Mannheim.
7. Krauss,G. (1997) Biochemie der Regulation und Signaltransduktion. Wiley-VCH, Weinheim.
8. Montminy,M. (1997) Transcriptional regulation by cyclic AMP. Annu. Rev. Biochem. 66, 807-22.
9. Takai-Igarashi,T. and Kaminuma,T. (1998) A Pathway Finding System for the Cell Signaling Networks Database. In Silico Biol., 1, 0012. www.bioinfo.de/isb/1998/01/0012/
10. POET Software Corporation (1999) POET 5.1 Object server & Java SDK, San Mateo, CA. <www.poet.com>
11. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res., **28**, 316-319.

GENOMICS-AIDED DRUG DEVELOPMENT: POPULATION GENOMICS AND INFORMATICS AT WORK

Richard Sidney Judson

Genaissance Pharmaceuticals, Inc. 5 Science Park, New Haven, Connecticut 06511, USA
e-mail: r.judson@genaissance.com

Keywords: development, drug, genomics

Resume

Fields ranging from architecture to manufacturing have been aided by computer simulation and visualization tools that have replaced prototypes and physical models. In much the same way, Genaissance has developed its population genomics and informatics capabilities to enable Genomics-Aided Drug Development (GADD). GADD is an integrative process that leverages pharmaceutical companies' past efforts in clinical trials into the future advancement of drug franchises. GADD results in increases in treatment efficacy (treatment delta) and in a reduction in the standard deviation of the response profile of the trial. The net effect of GADD is smart trials with substantially reduced cohort size. GADD is possible through Genaissance's proprietary HAP™ Technology. HAP™ Technology integrates HAP™ Markers (from haplotypes), the genetic markers of highest resolution, with our *DecoGen*™ Informatics Platform software. HAP™ Technology extracts information from genetic markers predictive of clinical outcome from retrospective trials or selected patient cohorts numbering no more than 200 patients. The utilization of these genetic markers in GADD results in the enrichment of trials for responders and the exclusion of non-responders.

HAP™ Technology is unique in its capability to capture genetic signals from cohorts consistent with the size of typical Phase I or Phase II trials or from review of selected patients in Phase III and Phase IV trials. An illustrative example of our HAP™ Technology at work is the successful discovery of beta-2 adrenergic receptor HAP™ Markers that are highly predictive of response or non-response to albuterol. In this study, HAP™ Technology classified 20% of the examined patients in a 120-patient trial as unequivocal non-responders and 15% as extreme responders. When this data is applied to GADD, the enrichment of response is dramatic. This lecture will describe HAP™ Technology and the various scenarios for its application to GADD. Various statistical and informatics tools will also be demonstrated by way of our *DecoGen*™ Informatics Platform.

We envision GADD will empower pharmaceutical companies in their strategic decision making. Specifically, GADD offers a new window on the utilization of genetics at the critical juncture between Phase II and Phase III for enabling or discontinuing drug development in larger human cohorts.

KNOWLEDGE BASE ON MOLECULAR-GENETICAL FOUNDATIONS OF LIPID METABOLISM REGULATION: CURRENT STATE AND PERSPECTIVE

Ignatieva E.V., Likhoshvai V.A., Ratushny A.V., Kosarev P.S.

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

Keywords: knowledge base, database, lipid metabolism

Resume

Motivation:

Development of computer resources including both databases and computer programs, together with the novel knowledge about the objects, obtained by theoretical analysis, is very promising for in silico prediction of behavior of live systems.

Results:

The computer system on molecular-genetical bases of lipid metabolism regulation contains the following modules: a) section of TRRD database on lipid metabolism genes regulation (LM-TRRD); b) samples of 5' flanking gene regions and transcription factor binding sites represented in the format of the SAMPLES database; c) section of GeneNet database describing sub-system on intracellular cholesterol regulation; d) mathematical model on intracellular cholesterol level regulation.

Availability:

http://wwwmgs.bionet.nsc.ru/mgs/dbases/knowlipidbase/

Introduction

One of the examples of intricate systems providing vitally important organism function is the system of lipid metabolism. Lipids refer to the important class of complex molecules acting in animal cells and tissues. The diversity and lipid level in a cell, tissue, and organ is determined by lipid metabolism (LM) processes including transport of lipids, their consumption and intracellular utilisation, de novo synthesis, degradation, and excretion. The processes of lipid metabolism require the involvement of numerous proteins with different functions. These proteins together with genes encoding such proteins are the components of LM system. Over the past years, the bulk of experimental data on lipid metabolism genes transcription regulation grows rapidly. An interest for the LM system is extremely high due to the facts that distortions in lipid metabolism are the reason of a seria of heavy human diseases including atherosclerosis, ischemia, etc. [Schmitz,G. et al., 1998; Chamberlain,J.C. and Galton D.J., 1990; Karpe,F. 1997, Tunstall-Pedoe H. and Smith W.C.S., 1990]. Besides, LM system is an example of well-studied system supporting homeostasis of important cell components including cholesterol. Previously we have reported about the section of the TRRD database accumulating the knowledge on lipid metabolism genes transcription regulation and about the results of theoretical analysis of information from this section [Ignatieva E.V. et al., 1997]. The goal of the present work is to systematize, generalize, and further analysis of the data on molecular-genetic bases of lipid metabolism regulation by means of computer system including databases, mathematical model, and computer programs.

Methods

Data on gene transcription regulation were accumulated in TRRD database [Kolchanov N.A. et al., 2000]. Besides, information on primary sequences of 5' flanking regions and transcription factor binding sites sequences were systematized in the format of SAMPLES database [Vorobiev D.G., 1998], by using original program [Kosarev P.S., 2000]. The data on the mechanisms of cholesterol level regulation in a cell was stored in the GeneNet database [Kolpakov, F.A., et al., 1998], by interactive data input via the Internet [Kolpakov F.A., Ananko E.A., 1999].

Current content of the knowledge base

***LM-TRRD* database.** A section of TRRD database on lipid metabolism gene transcription regulation, LM-TRRD, was developed. It includes information about 75 genes. In dependence on the function of the proteins encoded, the genes represented in the database may be divided into six groups: 1) genes encoding proteins involved in lipid transport; 2) genes of enzymes involved in lipid biosynthesis; 3) genes of enzymes involved in

lipid degradation; 4) genes of cell surface receptors; 5) genes of hormones, 6) genes of transcription factors involved in regulation of genes referring to the first five groups. Involvement of the genes of the last group in LM-TRRD is necessary for complete understanding of LM system regulation. The data on 118 regulatory regions and 457 transcription factor binding sites are collected in the section. This information was extracted by annotating of 284 scientific publications.

According to LM-TRRD data, more than 40 various transcription factors participate in lipid metabolism gene expression regulation. Moreover, the binding sites of some types are represented in regulatory regions of the gene group considered rather frequently. In particular, an information on 26 C/EBP binding sites, 19 COUP-TF binding sites, 17 SREBP-1 binding sites, 16 HNF-4 binding sites, 15 PPAR/RXR binding sites, 14 Sp1 binding sites is accumulated in the LM-TRRD section. The number of experimentally studied binding sites of these and some other types represented in the LM-TRRD section at least 4 times, is shown in Fig. 1 by dark columns. Following the hypothesis on the evenly distribution of binding sites within regulatory regions of all the genes described in TRRD database, we have calculated the expected number of sites of each of the types discussed above. An expected number of sites analyzed is demonstrated in Fig. 1 by columns with patchwork. For the most site types (except the sites like Sp-1 and Ap-1), the real number of sites represented in the LM-TRRD exceeds the expected values. By comparing real and expected site numbers according to X^2 criterion, the significant differences were found for a seria of sites (C/EBP, COUP-TF, SREBP, HNF-4, PPAR/RXR, HNF-1, RAR/RXR). Although the expected value for the Sp1 binding site, represented 14 times in the LM-TRRD, was higher than the real one, the statistically reliable discrepancies were not found. This result is in a good agreement with the evidence that this transcription factor is distributed ubiquitously and it regulates an expression of a wide variety of genes with different functions.

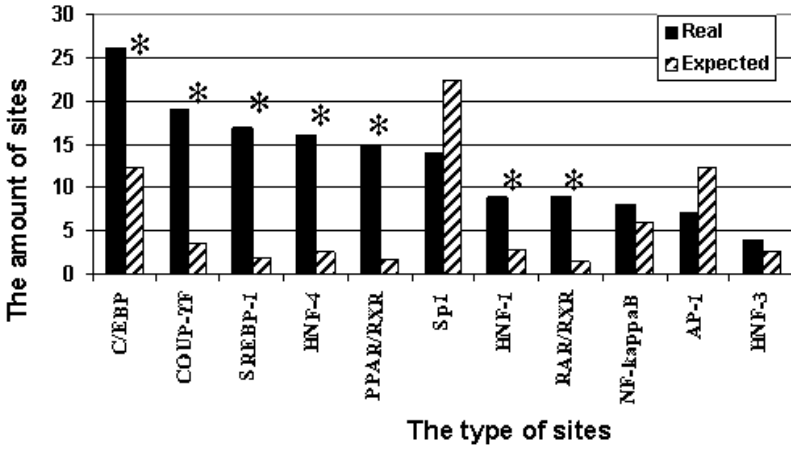


Figure 1. Number of transcription factor binding sites in regulatory regions described in LM-TRRD. Significant ($p<0.005$) differences between the number of real and expected sites of the fixed type are noted by asterisks.

SAMPLES database. Basing on information accumulated in LM-TRRD, we have compiled the following samples: 1) containing the sequences of 5' regulatory regions of genes involved in lipid metabolism (in-between -400/+50 and -200/+50 positions relatively transcription start site); 2) sites binding the following transcription factors: SREBP, PPAR/RXR, COUP-TF, C/EBP, HNF-4, and RAR/RXR. The samples of sites are represented by the fragments of regulatory regions with the length of 300 and 120 nucleotides, including an experimentally characterized site of a definite type given in the central position. Informational content of the samples from the SAMPLES database is described in the Table 1. It should be noted that the sites of COUP-TF, C/EBP, HNF-4, and RAR/RXR types are present in the regulatory regions of genes with a wide variety of functions, so the samples of these sites are formed using the data from all the TRRD sections.

Table 1. Informational content of the SAMPLES database.

Name of a sample	Content	Number of entries
LipMet_250	5'flanking regions of genes in the LM system	55
LipMet_450	within the regions in-between -200/+50 and -400/+50 positions, respectively	45
SRE_120_2000	SREs (sterol regulatory elements)	20
SRE_300_2000	interacting with SREBP factors	16
PPRE_120_2000	PPRE (peroxisome proliferator responsive element)	15
PPRE_300_2000	interacting with PPAR/RXR factors	12
COUP-TF_120_2000	COUP-TF sites interacting with factors COUPalpha (=Ear3/COUP-TF= EAR3)	30
COUP-TF_300_2000	and COUPbeta(= ARP-1=Ear-2), along with heterodimers of these factors and RXR factors.	25
C/EBP_120_2000	C/EBP sites binding to	25
C/EBP_300_2000	C/EBPalpha, C/EBPbeta, and C/EBPdelta factors	20
HNF-4_120_2000	HNF-4 sites binding to HNF-4 factor	25
HNF-4_300_2000		20
RARE_120_2000	RARE (retinoic acid-responsive element) interacting	12
RARE_300_2000	with RAR factors and RAR/RXR heterodimers	10

GeneNet database. At the diagram "Cholesterol" of the GeneNet database, there is a model of cholesterol regulation in a cell. The central elements of regulation are SREBP transcription factors. They are formed out of the molecule-precursor (preSREBP) under the action of sterol-regulated protease (SRP). In its turn, SRP activity is suppressed by high cholesterol level in a cell. Inner cellular cholesterol content is controlled by the mechanism with negative feedback. Under the low cholesterol level, the preSREBP transforms into the active form, SREBP. These proteins activate transcription of genes, controlling two cellular processes - cholesterol biosynthesis and cholesterol transport from the extracellular space. The SREBP activated genes, controlling cholesterol biosynthesis are: 3-hydroxy-3-methylglutaryl-CoA synthase, 3-hydroxy-3-methylglutaryl-CoA reductase, farnesyl diphosphate synthase, squalene synthase. The SREBP activated gene, controlling cholesterol uptake is the low-density lipoprotein receptor (LDLR) gene. Activation of LDLR expression enhances penetration into the cell of low-density lipoproteins (LDL), which are responsible for penetration of cholesterol, incorporated in these particles, into a cell. An increase in cholesterol level suppresses activity of sterol-regulated proteases. This, in turn, decreases both cholesterol biosynthesis and its penetration form outside.

Modeling. Modeling of dynamics of functioning of molecular-genetic system on cholesterol regulation is made on the basis of a model described in this issue [Ratushny A.V. et al., 2000]. This mathematical model may calculate the quantitative parameters of the gene network supporting cholesterol level in a cell. These parameters take into account different impacts: mutations, external stimuli, artificial alterations of a system, including those at the level of regulatory mechanisms, etc. For an example, Fig. 2 illustrates kinetics of inner cellular cholesterol in a normal cell and in a cell with mutation causing a 30% decrease in receptor production rate. At that, equilibrium concentration of LDL particles in blood is increased approximately two-fold. The cell reaction in response to increase of low-density lipoproteins in the outer cellular space for ~300 mg/dl becomes less expressed. Exogenous cholesterol penetration into the cell falls due to the shortage of LDL receptors. The mutant system returns to equilibrium concentration of LDL in blood rather slowly. The similar pattern is observed under homozygous form of hypercholesterolemia. This disease is characterized by the absence of LDL receptors in disease carriers, whereas high LDL concentration in blood is a risk factor for atherosclerosis [Goldstein,J.L. and Brown,M.S. 1989, LaCharity,L.A., 1998].

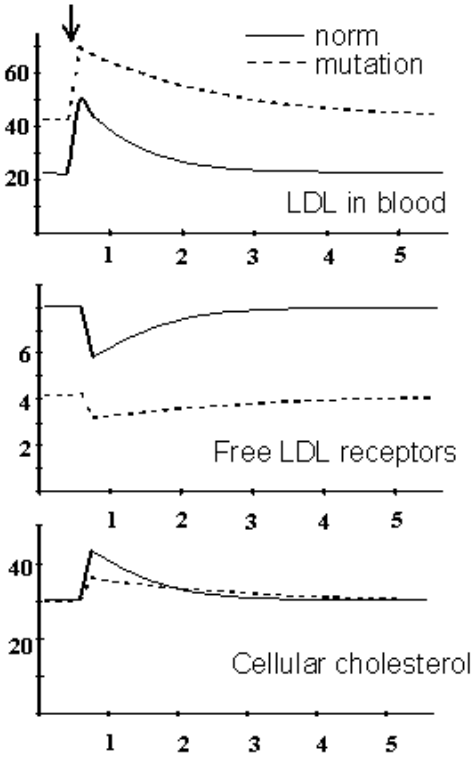


Figure 2. Normal and mutant cells responses to increase in LDL concentration up to 300 mg/dl in 0,5 hour after beginning of an experiment. By X-axis, time in hours, by Y-axis, concentration in 10^{+4} particles per cell.

Further development of the knowledge base

In future, we plan further development of the knowledge base on molecular-genetic foundations of lipid metabolism regulation. The current databases will be essentially supplemented. Besides, the following novel modules will be included into the knowledge base:

1. Knowledge base on transcription factor binding sites, containing the following sections:

- the data of computer analysis of the SAMPLES database and enabling to perform site recognition;
- consensuses and weight matrices of the sites;
- information on physico-chemical properties of sites obtained in result of computer data treatment by the system ACTIVITY [Ponomarenko M.P. et al., 1999]

2. Software programs for site recognition developed on the basis of approaches implemented in ACTIVITY system [Ponomarenko M.P. et al., 1999]

3. Knowledge on potential transcription factor binding sites in regulatory regions of genes involved in lipid metabolism. These potential sites will be predicted by recognition programs documented in the GeneExpress system [Kolchanov,N.A. et al., 1999].

4. Database on quantitative parameters used for modelling of cholesterol level regulatory system (constants of enzymes, concentrations of main proteins and low molecular components, etc.).

5. An interface for studying dynamics of molecular-genetic system on cholesterol synthesis regulation by using mathematical model, described in details in this issue [Ratushny A.V., et al., 2000].

Acknowledgements

The work was supported by the Russian Foundation for Basic research (grants Nos 98-04-49479, 98-07-91078, 99-07-90203, 00-04-49229, 00-04-49255, 00-07-90337), the US DOE grant DE-FG02-00ER62893/535228, and Integration Project of SB RAS No 66. The authors are grateful to I.V. Lokhova for the help with literature sources, to D.G. Vorobiev – for installation of site samples in the Internet, to G.Orlova – for help in translation of the manuscript.

References

1. Chamberlain J.C., Galton D.J. (1990) Genetic susceptibility to atherosclerosis. *British Medical Bulletin.* 46, 917-940.
2. Ignatieva E.V., Merkulova T.I., Vishnevsyi O.V., Kel A.E. (1997) Transcription regulation of lipid metabolism genes as described in the TRRD database. *Mol.Biol.(Mosk),* 31, 575-591.
3. Goldstein,J.L., Brown,M.S. (1989) Familial hypercholesterolemia. In Server C.R., Bendit A.L., Sly W.S., Valle D. (eds). *The Metabolic Basis of Inherited Disorders - 6th edition. McGraw Hill, New York*, 1215-1250.
4. Karpe F. Postprandial lipid metabolism in relation to coronary heart disease. (1997) *Proc. Nutr. Soc.* 56, 671-678.
5. Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N., Korostishevskaya,I.M., Romashchenko,A.G., Overton,G.C. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Res.,* , 28, 1, p. 298-301.
6. Kolchanov N.A., Ponomarenko M.P., Frolov A.S., Ananko E.A., Kolpakov F.A., Ignatieva E.V., Podkolodnaya O.A., Goryachkovskaya T.N., Stepanenko I.L., Merkulova T.I., Babenko V.N., Ponomarenko J.V., Kochetov A.V., Podkolodny N.L., Vorobyev D.G., Lavrushev S.V., Grigorovich D.A., Kondrakhin Yu.V., Milanesi L., Wingender E., Solovyev V.V., and Overton G.C. (1999) Integrated databases and computer systems for studying eukaryotic gene expression. *Bioinformatics*, 15, 669-686.
7. Kolpakov,F.A., Ananko,E.A. Kolesov,G.B. Kolchanov,N.A. (1998) GeneNet: a gene network database and its automated visualization. *Bioinformatics,* 14, 529 – 537.
8. Kolpakov F.A., Ananko E.A. (1999) Interactive data input into the GeneNet database. *Bioinformatics,* 15, 713-714.
9. LaCharity,L.A. (1998) Genetic risk factors in the development of heart disease: familial hypercholesterolemia and hyperhomocysteinemia. *AACN Clin Issues* 9, 531-538.
10. Kosarev P.S. (2000) TRRDEXTR: computer program for extraction of regulatory sequences described in TRRD database Proceedings of BGRS2000, Novosibirsk, *this issue.*
11. Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Podkolodny N.L., Savinkova L.K., Kolchanov N.A., and Overton G.C. (1999) Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins. *Bioinformatics*, 15, 687-703.
12. Ratushny A.V., Likhoshvai V.A., Ignatieva E.V., Matushkin Yu.G. (2000) Mathematical model of cholesterol regulation in a cell. Proceedings of BGRS2000, Novosibirsk, *this issue.*
13. Schmitz G., Aslanidis C., Lackner K.J. (1998) Recent Advances in Molecular Genetics of Cardiovascular Disorders - Implications for Atherosclerosis and Diseases of Cellular Lipid Metabolism. *Pathol. Oncol. Res.* 4, 153-161.
14. Tunstall-Pedoe H., Smith W.C.S. (1990) Cholesterol as a risk factor for coronary heart disease. *British Medical Bulletin*, 46, 4, 1075-1087.
15. Vorobiev,D.G., Ponomarenko,J.V., Podkolodnaya,O.A. (1998) Samples And Aligned: Databases For Functional Site Sequences. *Proc. I Intern. Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Russia,* 1, 58-61.

ACTIVITY: A DATABASE ON DNA REGULATORY SITES ACTIVITY, ADAPTED FOR ANALYSIS OF DNA-PROTEIN INTERACTIONS

Ponomarenko J.V., Furman D.P., Ponomarenko M.P., Orlova G.V., Frolov A.S., Podkolodny N.L., [1]Sarai A.

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

Keywords: activity, test-system, DNA/protein binding, hybrid model

Motivation:

Following experimental data, the estimations of relationships "site⇒pattern", which were obtained by different methods, better correlate to each other than to site activity, since various types of activities and even analogous activities measured in different test-systems may be not correlated at all.

Results:

The database ACTIVITY on DNA sites activities was supplemented by subsections SYSTEM (test-systems for measuring activity) and MODEL (models of application of relationships sequence⇒activity from one test-system to the others). Novel possibilities of such hybrid models have been demonstrated and discussed by the example of the YY1-site.

Availability:

ACTIVITY database, URL=http://wwwmgs.bionet.nsc.ru/systems/activity/.

Introduction

As follows from “mosaic” concept ("jigsaw puzzle") [1], transcription machinery consists from protein/protein-interactions between transcription factors and DNA/protein-interactions between these factors and DNA sites. The methods designed for recognition of DNA sites account only the "clue/lock" principle of DNA/protein-binding statistical mechanics. In accordance with this principle, the values of biological activity, DNA/protein-affinity, and relationships "site⇒pattern" correlate to each other [2]. Due to experiments, evaluations of relationships "site⇒pattern", obtained by different methods, better correlate to each other than to site activities [3], whereas different activity types and even one-type activities in different test-systems may be not correlated at all [4, 5]. In [6], a description is given of ACTIVITY database on site activities and relationships "sequence⇒activity". In the present paper, we have supplemented the ACTIVITY database by subsections: SYSTEM (test-systems for measuring activity) and MODEL (models of application of relationships sequence⇒activity from one test-system to the others). Novel possibilities of application of such models are demonstrated by the example of the YY1-site.

Materials and Methods

The ACTIVITY database [6] has four main subsections: ACTIVITY, experimental data; KNOWLEDGE, relationships "sequence⇒activity"; SYSTEM, test-system; MODEL, application of relationships "sequence⇒activity" from one test-system to the others. In Fig.1a, an example of the document stored in subsection ACTIVITY is given (fields **MI**, **MN** refer to the name of a document; **SN**, **OG**, **OS**, **AU**, **PN** - to description of activity; **SC**, **SA**, **PA –** experimental data; **DR**, **WW**, Web-resources). This subsection is addressed to the data analysis. The results of analysis are stored in the KNOWLEDGE database (Fig.1b: **CF**, **AB**, **PV**, description of relationships "sequence⇒activity"; **LC**, coefficient of linear correlation; **C-**, the program for prediction of activity for a test-system indicated in the field "**DR SYSTEM**"). An example of the document stored in SYSTEM subsection is given in Fig.1c (**EX**, motivation of the authors of an experiment; **RE**, conclusion made by authors of an experiment; **DT**, experimental conditions). This subsection accumulates interpretations of predictions. An example of the document from MODEL subsection is shown in Fig.1d (**MD**, the type of hybrid model; **LC**, **XI**, **ST**, **NT**, statistical estimations of reliability of application the relationships "sequence⇒activity" from one test-system to the others). This very subsection gives statistical verification necessary for application of relationships from one test-system to the others. This is the novelty of our approach.

Results and Discussion

In [6], a detailed description is given of DNA/activity-relationship between DNA $\{s_{-89}...s_is_{i+1}...s_{-66}\}$ of YY1-site and YY1-repressed transcription of luciferase gene inserted into pTiLUC plasmid with the minimal promoter ({AdML TATA} + {SV40E Inr} + {YY1-site in the region [-89; -66]}) in HeLa cells (Fig.2a):

$$Ln[luc_{pTiLUC,\ HeLa,\ [-79;-67]}] = -142.41+4.08\times[\Sigma_{-79\le i\le -68}\ \Omega(s_is_{i+1})/11], \quad (1)$$

:where Ln is a natural logarithm; $\phi\varphi$ - dinucleotide; $\Omega(\phi\varphi)$ – Twist angle value (^{O}) [7].

For YY1-repression of promoter pGL2 by displaced from the optimum position of YY1-site [4], by the equation (1), we have found the model "sequence⇒activity" (Fig.2b):

$$Ln[luc_{pGL2,\ HeLa}] = -1.30+0.10\times Ln[luc_1], \quad (2)$$

where $Ln[luc_1]=Ln[luc_{pTiLUC,\ HeLa,\ [-3;9]}]$, equation (1); [-3; 9] from the optimal YY1-site [4].

For YY1-repression of luciferase gene transcription in plasmid pTiLUC in the cells PYS-2 [4] by the equation (1), a model "sequence⇒activity" was detected (Fig.2c):

$$Ln[luc_{pTiLUC,\ PYS\text{-}2}] = 1.71+3.12\times Ln[luc_2], \quad (3)$$

where $Ln[luc_2]=Ln[luc_{pTiLUC,HeLa,[-79;-67]}]-M_{-89;-66}Ln[luc_{pTiLUC,HeLa}]$; $M_{-89;-66}$, averaging of the equation (1).

Equation (3) indicates that in PYS-2 cells under the presence of YY1-site, biologically active was the region [-79; -67] independently of localization of this YY1-site, besides, local surroundings of this active region [-79; -67] could decrease YY1-repression by competition with this region for the binding of the YY1 factor. This is, indeed, the novelty.

For YY1/DNA-affinity [4], by equation (1), a hybrid model was derived (Fig.2d):

$$Ln[YY1/oligoDNA\text{-}affinity] = 5.32-1.27\times Ln[luc_3], \quad (4)$$

where $Ln[luc_3]=M_{-11;3}(Ln[luc_{pTiLUC,HeLa}])$, averaging of equation (1), [-11; 3] from "CAT" YY1-site.

In Fig.1d one can see the folllowing: top, input of equation (4) into MODEL subsection; bottom, input of the model of equation (1) application to recognition of YY1-affine DNA [8] from random ones(Fig.2e):

$$\Phi[YY1_{SELEX}/Rand] = -2.45\times(3.05+Ln[luc_4]), \quad (5)$$

where $Ln[luc_4]=MAX_{-7;6}Ln[luc_{pTiLUC,HeLa}]$, maximum of equation (1); [-7; 6] from "CAT" YY1-site.

Due to equation (5), SELEX-protocol [8] selected DNA with the highest values of minimal YY1-affinity in randomization region of synthesized DNA ("Maximin" principle). This is one more novelty. For recognition of experimentally known natural YY1 sites [9] out of random DNA by equation (1), we have derived a hybrid model (Fig.2f):

$$\Phi[YY1_{NATURAL}/Rand] = -1.11\times(4.06+Ln[luc_5]), \quad (6)$$

where $Ln[luc_5]=Ln[luc_{pTiLUC,\ HeLa,\ [-9;3]}]$, equation (1); oligoDNA region [-9; 3] from "CAT" of YY1-site.

The difference between the models of YY1-site (equation 6) and YY1-affine DNA (equation 5) corresponds to the difference between their consensuses [8]. For YY1-affinity of mutant YY1 sites in promoter TdT in mouse [5], by equation (1), there was found the hybrid model (Fig.2g):

$$Ln[YY1/DNA_{MUTANT}\text{-}affinity] = -10.94-2.26\times Ln[luc_5]. \quad (7)$$

Though equation (1) predicts YY1-repression of transcription, this equation became a basis for a hybrid model for YY1-activation of transcription [5] (Fig.2h):

$$Ln[Inr] = -6.00-1.27\times Ln[luc_5], \quad (8)$$

where $Ln[luc_5]=Ln[luc_{pTiLUC,\ HeLa,\ [-9;3]}]$, equation (1); [-9; 3] relatively transcription start.

Equations (6, 7, 8) indicate {transcription start}⇔{YY1-site of activation of transcription}. For YY1/Inr-affinity [5], the hybrid model of application of equation (1) was as follows (Fig.2i):

$$Rank[YY1/Inr\text{-}affinity] = -9.38-2.15\times Ln[luc_6], \quad (9)$$

where $Ln[luc_6]=Ln[luc_{pTiLUC,HeLa,[-10;2]}]-Ln[luc_{pTiLUC,HeLa,[-9;3]}]+Ln[luc_{pTiLUC,HeLa,[-8;4]}]$; for equation (1), DNA regions are given relatively transcription start; Rank - range determined in [5].

Following equations (6-9), positive impact into YY1/Inr-affinity is made at positions -1 and +2 from transcription start bringing the negative effect, these facts being in coincidence with experimentally found in [5] discrepancy of consensuses of YY1-site (C C A T) and Inr-element (C/t C/t A N T/A c/T c/t): insertions {"N" between similar parts "CCA" and "T"} or {"AN" between "CC" and "AT"}.

Considered hybrid models of YY1-dependable biological activities (equations 2-9) relying on the basal model (equation 1) are documented in MODEL subsection (Fig.1d). The search and documentation of such models ia a novelty of our approach.

All the models "sequence⇒activity" are in accordance with "mosaic" concept ("jigsaw puzzle") [1], which supplements DNA/protein-interactions "clue-lock" by accounting protein/protein interactions. Following equations (1-3, 6-8), protein/protein interactions may shift biologically active region of DNA/protein-binding relatively its consensus. Since DNA-protein affinity of such "mosaic"-dependable site is weakened by this shift in this case, local surroundings may compete with this site for the protein binding, thus providing a mimicry (equations 4 and 5) and inhibition (equations 3 and 9) of biological activity. Hence, models of application of one test-system regularities to others give novel knowledge on DNA/protein- and protein/protein-interactions.

The present study was supported by Russian Foundation for Basic Research (98-07-90126) and STA Fellow #499042 (Japan).

a)
```
MI  A00J0006
MN  Transcriptional activity YY1-repressed
DR  KNOWLEDGE: K00J0006
DR  SYSTEM: T00J0006
SN  Transcription factor YY1 binding site
OG  Synthetic oligoDNAs of YY1-high-affinity
OS  Insertion into plasmid transfected HeLa cell
AU  % relative to the vector without YY1-sites
PN  Transcription start
DR  REFERENCES: RF0J0004
WW  FIGURE (DNA-protein): http://…/yy1x.html
WW  FIGURE (Data source): http://…/yy1d.html
SC  #11, YY1-core, "CAT", on the (-) chain
    TCGTTAGGAC TTAAAATGGC GTC
SA  12
PA  89
    . . . . . .
SC  #79, YY1-core, "CAT", on the (+) chain
    GTCGTCCATA TTGTAATGGC GTC
SA  97
    . . . . . .
SC  #14, none YY1-core, "CAT"
    TCGTTTAGTT AATACTTCGC GTC
SA  97
//
```

b)
```
MI  K00J0006
MN  Transcriptional activity YY1-repressed
DR  SCIENTISTS: SCI00001
DR  ACTIVITY: A00J0006
DR  MODEL: M00J0006
DR  SYSTEM: T00J0006
WW  TOOLS: http://www…/YY1Repr.html
CF  conformation property means, region [A;B]
DR  PROPERTY: P0000015
PV  Twist (DNA-protein complex)
AB  1 12; relative to -80 of minimal promoter
LC  -0.803
WW  FIGURE: http://www…/YY1ReprInc.html
C-  double TwDnaProt_for_YY1Repr (…){
    double DinucPar[16]={
    //_AA__AT__AG__AC__TA__TT__TG__
      35.6, 29.3, 31.9, 31.1, 39.5, 35.6, 36.0,
    . . . . . .
    return (X/(double)(SiteLength-1));}
CF  Prediction activity by simple regression
AB  1 12; relative to -80 of minimal promoter
C-  double YY1_Repr_by_TwDnaProt (…){
    . . . . . .
    return (-142.41+4.0819*x1);}
//
```

c)
```
MI  T00J0006
MN  Transcriptional activity YY1-repressed
DR  REFERENCES: RF0J0004
EX  To determine if the YY1-site can affect
EX  transcription they were cloned upstream
EX  minimal promoter controlling reporter gene.
DT  Reporter = luciferase gene (LUC-activity).
DT  Promoter = AdML TATA and SV40E Inr
DT  YY1-site = synthetic oligoDNA randomized/
DT  selected by YY1-affinity (SELEX-protocol).
DT  YY1-site insertion into the region [-80; -70].
DT  Either (-)/(+) orientation of YY1-core, "Cat".
DT  Plasmid pTiLUC construct was transfected
DT  into HeLa cells to measure LUC-activity
DT  YY1-repressed in transcription initiation.
DT  Luciferase activity of the plasmid construct
DT  pTiLUC without YY1-core CAT in-between
DT  -89 and -66 positions was labeled "100%".
DR  High YY1-affinity = Low LUC-activity.
WW  FIGURE-cited: http://wwwmgs…/yy1d.html
DR  ACTIVITY: A00J0006
RE  YY1 can repress transcription regardless
RE  orientation from a synthetic basic promoter
DR  SELEX_DB: S00J0031
DR  ACTIVITY: A00J0007
DR  ACTIVITY: A00J0006a
DR  ACTIVITY: A00J0005
DR  MODEL: M00J0006
//
```

d)
```
MI  M00J0006
MN  Transcriptional activity YY1-repressed
DR  ACTIVITY: A00J0006
DR  SYSTEMS: T00J0006
WW  TOOLS: http://www…/YY1Repr.html
CF  Activity adapted to other test-system
DR  ACTIVITY: A00J0005
DR  TEST-SYSTEMS: T00J0005
MN  YY1-affinity to synthetic oligoDNA's
MD  Mean activity
AB  -11 3, relative to YY1-site core "Cat"
LC  -0.689
WW  FIGURE: http://www…/j0006_j0005.html
C-  double YY1_aff_AJ0005_via_J0006 (…){
    . . . . . .
    return (-5.3242+1.2738*x1);}
CF  Activity adapted to site recognition
DR  SELEX_DB: S00J0008, S00J0031
MN  YY1-high-affinity DNA randomized/selected
MD  Maximal activity
AB  -7 6, relative to YY1-site core "Cat"
XI  93.45
ST  1.011 (2.491) 31.5%
NT  -0.987 (3.046) 39.0%
WW  FIGURE: http://ww…/j0006_yy1selex.html
C-  double YY1_SELEX_via_J0006 (…){
    . . . . . .
    if(x<x1)x=x1; x-=-3.05; x*=-2.45; return (x);}
//
```

Fig. 1. Examples of documents stored in ACTIVITY database, containing (a) experimental data on DNA sites activity (subsection ACTIVITY); (b) regularities "sequence-activity" (KNOWLEDGE), (c) description of test-systems (SYSTEM); (d) hybrid models for application of one test-system to the others (MODEL).

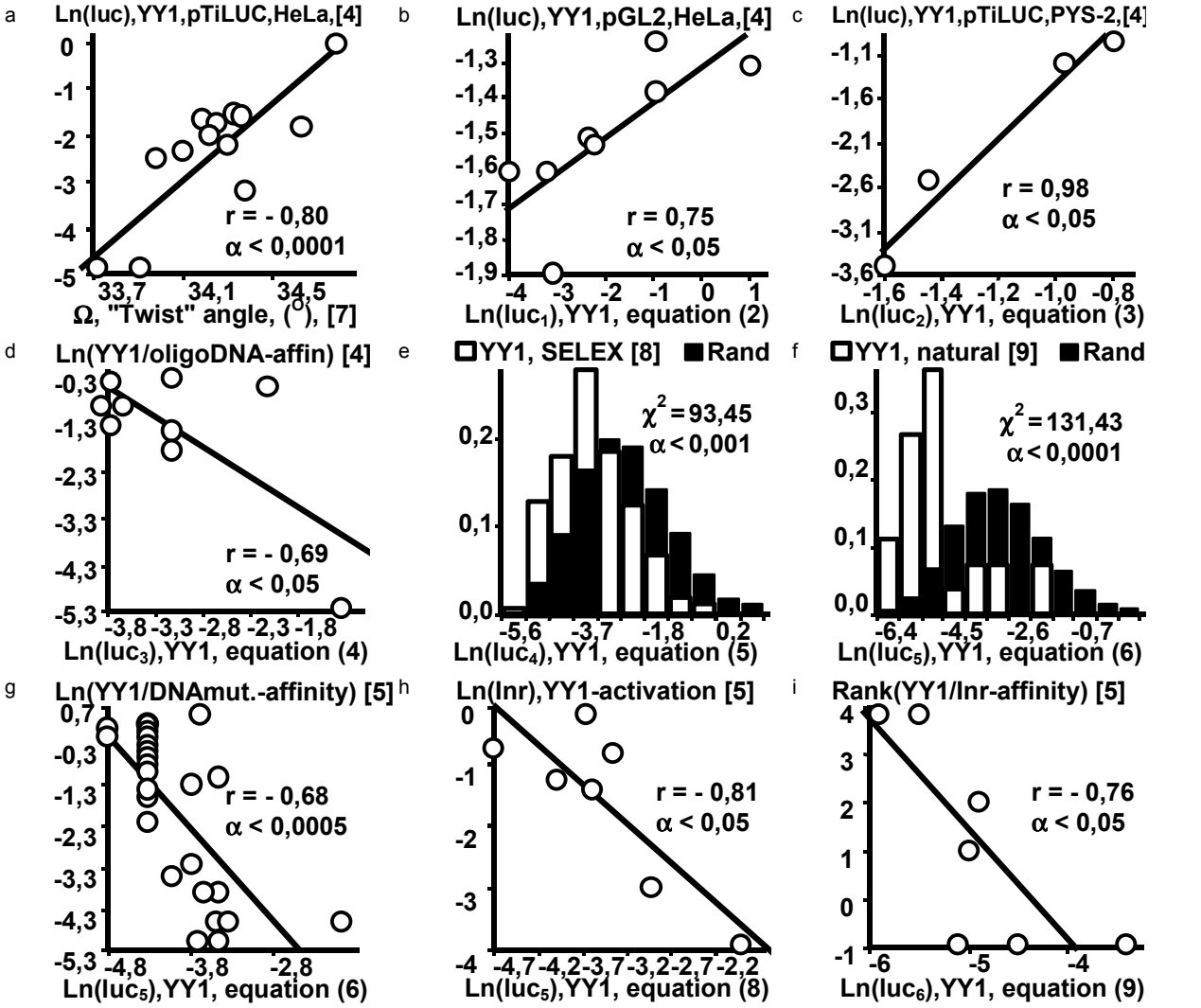


Fig.2. Basic (a) and hybrid (b-i) models of relationships "sequence⇒activity" for YY1-dependable biological activities documented in the MODEL subsection (Fig.1d).

References

1. Johnson,P. McKhight, S. (1989) Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.* **58**, 799-839.
2. Berg,O., von Hippel,P. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.,* **193**, 723-750.
3. Roulet,E., Fisch,I., Bucher,P. Mermod,N. (1998) Evaluation of computer tools for prediction of transcription factor binding sites on genomic DNA. *In Silico Biology*, **1**, 21-28.
4. Hyde-DeRuyscher,R., Jennings,E., Shenk,T. (1995) DNA binding sites for the transcriptional activator/repressor YY1. *Nucleic Acids Res.*, **23**, 4457-4465.
5. Javahery,R., Khachi,A., Lo,K., Zenzie-Gregory,B., Smale,S. (1994) DNA sequence require-ments for transcriptional initiator activity in mammalian cells. *Mol.Cell.Biol*, **14**, 116-127
6. Ponomarenko, M., Ponomarenko J., Frolov A., Podkolodny,N., Savinkova,L., Kolchanov,N., Overton, C. (1999) Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins. *Bioinformatics,* **15,** 687-703.
7. Suzuki,M., Yagi,N., Finch J. (1996) Role of base-backbone and base-base interactions in alternating DNA conformations. *FEBS Lett.* **397**, 148-152.
8. Ponomarenko,J., Orlova,G., Ponomarenko,M., Lavryushev,S., Frolov,A., Zybova,S., Kolchanov,N. (2000) SELEX_DB: an activated database on selected randomized DNA/RNA sequences addressed to genomic sequence annotation. *Nucleic Acids Res.,* **28**, 205-208.
9. Vasiliev,G., Merkulov,V., Kobzev,V., Merkulova,T., Ponomarenko,M., Kolchanov,N. (1999) Point mutations within 663-666 bp of intron 6 of the human TDO2 gene, associated with a number of psychiatric disorders, damage the YY1 transcription factor binding site. *FEBS Lett.,* **462**, 85-88.

RECOGNITION GROUPS: A NEW METHOD FOR DESCRIPTION AND PREDICTION OF TRANSCRIPTION FACTOR BINDING SITES

Kondrakhin Yu.V., [1]Milanesi L., Lavryushev S.V., [2]Schug J., Kolchanov N.A.

Institute of Cytology and Genetics of SB RAS, Novosibirsk, Russia
[1] Instituto di Tecnologie Biomediche Avanzate, Consiglio Nazionale Delle Ricerche, Milan, Italy
[2] Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA, USA
e-mail: kondrat@bionet.nsc.ru
*Corresponding author

Keywords: binding site, transcription factor, matrix approach, consensus, recognition accuracy

Resume

Motivation:

To try to increase the accuracy of transcription factor binding site prediction we propose a new approach of binding site description by means of recognition groups that are high-homology sets of short oligonucleotides. We perform an analysis of actual binding sites to identify and remove the non-significant flanking positions. To compare the accuracy of various recognition procedures we have developed the penalty function technique.

Results:

A new method for predicting eukaryotic transcription factor binding sites is described. A Recognition Group description of binding sites is proposed as an alternative to the commonly used consensus sequence or weight matrix representations. A Recognition Group (RG) is a high-homology family of oligonucleotides of the same length. Our algorithm does not need any parameter setup (http://wwwmgs.bionet.nsc.ru/ mgs/ programs/ yura/recgroup.html, http: //wwwmgs.bionet.nsc.ru/ mgs/programs/ yura/rgscan1.html); they are estimated automatically by a specially developed optimization procedure. The problem of comparing different recognition methods is also addressed. For this purpose we developed a penalty function technique allowing universal comparison of any recognition methods in terms of accuracy. Three methods for predicting transcription factor binding sites (consensus, weight matrix, and Recognition Group) have been analyzed. When applied to the scanning of long sequences, the newly developed recognition method appeared to be quite effective in terms of penalty function.

Introduction

When annotating eukaryotic genomes one often encounters problems identifying regulatory regions as well as specific transcription factor binding sites [1]. Information sources related to transcription factors and their binding sites can be found in the databases TRANSFAC [2,3], TRRD [4], COMPEL [5] and TFD [6]. These databases contain transcription factor binding site disposition and other structural and functional features obtained from experimental studies.

The original and most commonly used method of functional site description is based on the short IUPAC coded string called the *consensus sequence*. Various methods of consensus sequence construction have been proposed [7]. A significant number of consensus sequences built for specific binding sites are compiled in the databases mentioned above. Data compiled and described in [8,9] contain many consensus sequences. The consensus sequences in [9] were used for the analysis of regulatory regions and for the classification of promoters in eukaryotic DNA sequences [10].

Since 1990 [11] the weight (or nucleotide frequency) matrix approach has been applied to binding site description as an alternative to the consensus method. This approach, compared with a consensus sequence, retains more information about the site pattern. Matrices are commonly used for site description and are accumulated in databases such as TRANSFAC and IMD [12]. Recognition algorithms based on these matrices have been implemented in the computer programs such as MatInspector [13] and MATRIX SEARCH [12].

The task of constructing reliable and efficient methods for recognition of transcription factor binding sites is far from complete. In particular, presently available programs for recognition of potential transcription factor binding sites in long DNA sequences generally yield a huge amount of false positives [14].

Methods and algorithms

The construction of a recognition group $R = \{R_1,...,R_k\}$ for the binding site of a transcription factor is based on detailed analysis of a training set U consisting of m unaligned nucleotide sequences U_i, $i = 1,...,m$. These nucleotide sequences U_i have been extracted from the EMBL database using TRANSFAC database

references. In general, sequences U_i may be of different lengths and the actual binding site location inside U_i and strand orientation is unknown. The complete process of creating the recognition group consists of two parts: 1) the construction of recognition groups for a range of specific values of parameters *l* (binding site length) and T (the maximum admissible mismatches); and 2) the estimation of optimal values for parameters *l* and T.

At the initial stage of the algorithm the first oligonucleotide R_1 is found as the *l*-length oligonucleotide occurring most frequently in the expert set U. The rest of recognition group is determined by the iterative procedure. Each time a new oligonucleotide R_i is selected as the most frequent oligonucleotide, we reduce the residual training set U by excluding from it all the sequences containing this oligonucleotide R_i. The values for parameters *l* and T are estimated on the base of maximization of the function $Q_{T/}$

$$Q_{T/} = f_{T/} \times [\, (f_{T/} - f_{T\,/+1}\,) + (f_{T/} - f_{T-1\,/})\,],$$

where $f_{T/}$ is the relative number of sequences from the training set U containing the binding sites predicted using recognition group R. The group R, in turn, was built for given values of parameters T and *l*. In other words, the frequency $f_{T/}$ is the indicator of how well the actual binding sites are detected by recognition group constructed on the base of concrete parameters T, *l*. The closer $f_{T/}$ is to 1, the better the quality of the recognition group.

As rule, the accuracy of the recognition procedure has been checked by both type I error α_1 (false negatives rate) and type II error α_2 (false positives rate) *simultaneously*. For comparing two different recognition procedures it is convenient to represent the accuracy of each procedure with single generalized score instead of the pair (α_1,α_2). To do this we use a penalty function that allows us to treat each type of error differently. We penalize each error type according to its effect on the utility of predictions of binding sites in long (approximately 10^5 bp) sequences. Since the number of actual binding sites is small, we want to avoid the many potential false positives at the expense of a small percentage of false negatives. This leads us to penalize type II errors more heavily than type I errors. Thus, an accuracy of any recognition procedure with error rates α_1 and α_2 will be characterized by the penalty Ω which is defined as the sum

$$\Omega = \{\alpha_1 + 1 - \log(\alpha_2) / M, \text{ if } \alpha_2 > 10^M ; \alpha_1 \text{ if } \alpha_2 \le 10^M \},$$

where M(L) = -log(L) is pre-defined negative-valued parameter depending on the length L of target sequence scanned for (potential) binding sites by means of the recognition procedure considered. We suppose that recognition methods are applied for recognition of potential transcription factor binding sites in long sequences of length L ~ 10^5bp. Hence, parameter value M = -log(L) = -5 is considered.

Results and discussion

Twenty one training sets were compiled for the transcription factors such as AP1, AP2, ATF/CREB, C/EBP, COUP/RAR, GATA, MEF-2, OCT and Sp1. The sizes of all the training sets are greater than 8 sequences while the upper limit is 139 sequences for the Sp1 expert set. Table 1 contains four recognition groups for binding sites of transcription factors AP1, ATF/CREB, NF-Y and Sp1. Each oligonucleotide R_i in recognition group is accompanied by a *weight coefficient* w_i which reflects how often the oligonucleotide R_i occurs in training set. In particular, the NF-Y family consists of *k* = 10 oligonucleotides of length *l* = 7bp and the main motif R_1 = AGCCAAT has weight w_1 = 7. The remaining nine oligonucleotides R_i (i>1) differ from R_1 in one or two nucleotides. Therefore, the homology parameter T is equal to two mismatches, i.e., T = 2. The third oligonucleotide R_3 = AaCCAAT has single mismatch at the second position while the second oligonucleotide R_2 = gaCCAAT has two mismatches at the start. Mismatched positions are denoted by lowercase letters.

When the recognition group has been built, the matrix and the consensus are derived using the same training set after preliminary alignment. It is important to note that there is no need for the preliminary alignment of the training set and determination of site orientation in recognition group construction.

In the course of comparative analysis additional problems arise concerning the choice of parameter values for the matrix method. The threshold *crit* and site length *l* are nuisance parameters that have to be adjusted in each particular case. While investigating the influence of the parameters *crit* and *l* on the recognition accuracy we have observed the following effects. For fixed *crit* increasing length *l* leads to α_1 rising and α_2 falling. Once the length *l* achieves the certain value the type-I and type-II error rates stabilize because all (or almost all) conservative columns of matrix MAT are involved in the recognition process. The same effect is observed when considering the penalty function Ω instead of α_1 and α_2. Table 2 demonstrates the behavior of α_1 , α_2 and Ω as *l* increasing during the NF-Y binding sites recognition by the matrix method given threshold *crit* = 0.90. The accuracy is also sensitive to the value of the parameter *crit*. Table 3 obtained for NF-Y recognition demonstrates the effect common for all 21 binding sites analyzed: for fixed *l* = 15 bp the increase of *crit* threshold leads to α_1 rising and α_2 falling. Thus, considering type I and II errors, one would come to a contradiction: to minimize the α_1 the minimal admissible *crit* value should be used, whereas to minimize the α_2 the maximal *crit* threshold should be taken. Using the penalty function approach we can adjust *crit* so that the matrix recognition procedure would be as accurate as possible. As shown in the fourth column of Table 3, the penalty Ω has local minimum 0.691 for *crit* = 0.90.

Table 1. The recognition groups for binding sites of transcription factors AP1, ATF/CREB, NF-Y, and Sp1.

No	AP1 (*l* = 6bp; T = 1)		ATF/CREB (*l* = 8bp; T = 2)		NF-Y (*l* = 7bp; T = 2)		Sp1 (*l* = 6bp; T = 1)	
	RG	Weight	RG	Weight	RG	Weight	RG	Weight
1)	TGACTC	23	TGACGTCA	10	AGCCAAT	7	CCGCCC	79
2)	TGACTa	8	TGACGTag	3	GaCCAAT	4	CCtCCC	21
3)	TGACgC	5	TGACGTCt	2	AaCCAAT	3	CCaCCC	11
4)	TGACcC	4	TGACGTtt	2	AGCCAcT	1	CtGCCC	9
5)	TGAaTC	3	TGACGgaA	2	AaCCAAc	1	CCGCCt	7
6)	TGACTg	3	TGACGTgt	2	AGCCtcT	1	CCGgCC	3
7)	TcACTC	2	TGACGTaA	1	AGgCAAg	1	aCGCCC	3
8)	aGACTC	2	TGACaTCA	1	AtgCAAT	1	CaGCCC	2
9)	TaACTC	1	TGACGaCA	1	AGCCAtg	1		
10)	TGACTt	1	TGACGaaA	1	GGCaAAT	1		
11)			CGACaTCA	1				
12)			TGACGTgg	1				
13)			TGACGcaA	1				
14)			TGACGTac	1				

Table 2. Matrix method accuracy as a function of the site length *l* (*crit* = 0.90; factor = NF-Y).

l	α_1	α_2	Ω
5	0.227	0.010384	0.831
7	0.318	0.002028	0.780
9	0.318	0.001652	0.762
11	0.364	0.000568	0.714
13	0.409	0.000298	0.704
15	0.409	0.000258	0.691

Table 3. Matrix method accuracy as a function of *crit*. (site length *l* = 15bp; factor = NF-Y).

crit	α_1	α_2	Ω
0.70	0.136	0.044449	0.866
0.75	0.182	0.018980	0.837
0.80	0.227	0.006339	0.788
0.85	0.318	0.001790	0.769
0.90	0.409	0.000258	0.691
0.95	0.591	0.000037	0.705
0.999	1.000	$<10^{-5}$	1.000

Table 4 summarizes the results of the comparison between recognition methods applied to binding sites of transcription factors AP1, ATF/CREB, NF-Y, and Sp1. Note that the matrix method is represented by the best performance achieved by means of parameter adjustment. In general, comparing the Ω values of the recognition group approach to the matrix method it turn out that for all but 3 cases (ETF, IRF-1, SRF) the matrix method penalty values exceed ones for the recognition groups. In turn, the consensus sequences yield higher penalty values than that for the matrix method except for NF-kB. Thus, the developed technique of recognition groups seems to be effective in terms of the newly introduced penalty function. It incorporates a new method for binding sites description, leading to the results not worse than that for traditional weight matrix and consensus representation.

Table 4. The accuracy of binding sites recognition by means of recognition groups, matrix method and consensus.

Binding site	Recognition group			Matrix			Consensus		
	α_1	α_2	Ω	α_1	α_2	Ω	α_1	α_2	Ω
AP1	0.188	0.004303	0.714	0.156	0.007421	0.730	0.609	0.000221	0.878
ATF/CREB	0.147	0.000207	0.410	0.118	0.000964	0.514	0.520	$<10^{-5}$	0.529
NF-Y	0.045	0.001466	0.479	0.409	0.000258	0.691	0.591	0.000095	0.786
Sp1	0.029	0.008347	0.613	0.194	0.001999	0.654	0.647	0.000057	0.799

References

1. Wingender, E. (1994) Recognition of regulatory regions in genomic sequences. *J. of Biotechnology*, **35**, 273-280
2. Knuppel, R., Dietze, P., Lehnberg, W., Frech, K. and Wingender, E. (1994) TRANSFAC Retrieval program: a Network model database of eukaryotic transcription regulating sequences and proteins. *J. Comput. Biol.*, **1**, 191-198
3. Wingender, E., Dietze, P., Karas, H. and Knuppel, R. (1996) TRANSFAC: A database on transcription factors and their DNA binding sites. *Nucleic Acids Res.*, **24**, 238- 241.
4. Kel, A.E., Kel, O.V., Ischenko, I.V., Karas, H., Kolchanov, N.A., Sklenar, H. and Wingender, E. (1996) TRRD and COMPEL databases on transcription linked to TRANSFAC as a tool for analysis and recognition of regulatory sequences. *Proceedings of the German Conference on Bioinformatics, Leipzig,* 113-117.
5. Kel, O.V., Romaschenko, A.G., Kel, A.E., Wingender, E. and Kolchanov, N.A. (1995) A compilation of composite regulatory elements affecting transcription in vertebrates. *Nucleic Acids Res.*, **23**, 4097-4103.
6. Ghosh, D. (1992) TFD: the transcription factors database. *Nucleic Acids Res.*, **20**, 2091-2093.

7. Day, W.H.E. and McMorris, F.R. (1992) Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Res.*, **20**, 1093-1099.
8. Boulikas, T. (1994) A compilation and classification of DNA binding sites for protein transcription factors from vertebrates. *Crit. Rev. Euk. Gen Express*, **4**, 117-321
9. Faisst, S. and Meyer, S. (1992) Compilation of vertebrate-encoded transcription factors. *Nucleic Acids Res.*, **20**, 3-26.
10. Kondrakhin, Y.V., Kel, A.E., Kolchanov, N.A., Romashchenko, A.G. and Milanesi, L. (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Applic. Biosci.*, **11**, 477-488.
11. Bucher, P. (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, 212, 563-578.
12. Chen, Q.K., Hertz, G.Z. and Stormo, G.D. (1995) MATRIX SERCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput. Applic. Biosci.*, **11**, 563-566.
13. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.*, **23**, 4878-4884.
14. Pickert,L., Reuter,I., Klawonn,F. and Wingender,E. (1997) Transcription regulatory regions revealed by signal detection and fuzzy clustering. In Mewes,H.W. and Frishman,D. (eds.), *Computer Science and bioloqy: Proceedings of the German Conference on Bioinformatics, Bavaria,* pp.99-101.

MD-CAVE – THE METABOLIC DISEASES DATABASE A SYSTEM FOR STORING INFORMATION ABOUT HUMAN INBORN ERRORS

A. Freier, R. Hofestädt, M. Lange, U. Scholz and T. Töpel

Institute of Technical and Business Information Systems
Otto-von-Guericke-University, Magdeburg, Germany
e-mail: {freier|hofestae|mlange|uscholz|toepel}@iti.cs.uni-magdeburg.de

Resume

Due to the progress in molecular diagnostics the number of known inborn errors of metabolism increases rapidly. More patients with inborn errors are diagnosed and need a sufficient treatment [1]. Missing experience in treatment of seldom diseases, because of their heterogeneity and rarity, demands assistance of computer science. New developments - especially the facilities of the Internet - provide possibilities of implementations for inborn errors of metabolism in various forms. Aim of our research is to offer Internet-based systems for acquisition and representation of knowledge about inborn errors of metabolism.

The MD-Cave has been developed in order to enable a simultaneous coverage of molecular and clinical knowledge about metabolic diseases. Before MD-Cave, systems contained either microscopic (genetic, enzymatic, pathways) information or focussed on macroscopic clinical aspects. But especially for the selection of an appropriate therapy it is very important for the physician to gather information about the pathologic mechanism of a disease. Only if he knows which metabolic pathway is blocked, he can influence the accumulation of toxic metabolites with a diet or by activating alternative pathways.

Acknowledgments

This work is supported by the Federal Ministry for Education and Research (BMBF) in cause of the Human Genome Project and by the “Kurt-Eberhard-Bode-Stiftung im Stifterverband für die Deutsche Wissenschaft”.

Reference

1. H.L. Levy, Newborn screening by tandem mass spectrometry: a new era. *Clin Chem*, 44(12):2401-2402, 1998.

CONTEXTUAL FEATURES OF YEAST mRNA 5'UTRs POTENTIALLY IMPORTANT FOR THEIR TRANSLATIONAL ACTIVITY

*[1]Kochetov A.V., [1]Vorobiev D.G., [1,2]SirnikO.A., [3]Kisselev L.L., [1]Kolchanov N.A.

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2]Krasnoyarsk Technological University, Krasnoyarsk, Russia
[3]Engelhardt Institute of Molecular Biology, Moscow, Russia
e-mail: ak@bionet.nsc.ru
*Corresponding author

Keywords: mRNA, sequence characteristics, 5'UTRs, translation efficiency

Resume

Motivation:

It is well known that the efficiency of eukaryotic mRNA translation depends on its sequence characteristics. However, mRNA 5'UTR features influencing translation initiation efficiency are still poorly investigated. Here we describe 5'UTR characteristics potentially important for the efficient translation of yeast mRNAs.

Results:

General contextual features of yeast mRNA 5'UTRs were determined. Analysis of correlations of these characteristics with mRNA translational efficiency (evaluated by using the codon adaptation index, CAI) was used to reveal 5'UTR features influencing translatability. As was found, some 5'UTR features correlate with CAI significantly.

Introduction

It is known that translational efficiency of eukaryotic mRNAs varies with their sequence characteristics (for review, see Kozak, 1994; Futterer and Hohn, 1996; Pain, 1996). Sequence features of the mRNA 5' untranslated region (5'UTR) affect the rate of translation initiation and, thereby, the level of polypeptide production. It is widely accepted that the majority of eukaryotic mRNAs are translated through the linear scanning mechanism (for review, see Kozak, 1994). In the frames of this model, some features of the leader sequence affect mRNA translational efficiency (e.g., the context of AUG start codon, the presence of upstream AUGs within 5'UTR, and the stable hairpins within the leader). Apart from the context of the start codon, influencing its recognition efficiency, eukaryotic mRNA leader sequences are suggested to be highly divergent: none conservative translational signals were found (Pain, 1996). However, many experimental data proved the active role of 5'UTR sequences in translation initiation process. Translational enhancers are one of the most significant phenomenon of this sort: it was experimentally found that 5'UTRs of several plant viruses and cellular mRNAs are able to enhance translational efficiency of the heterologous downstream coding sequences. Mechanisms of their specific activity are not clear yet, however, one of the possible reasons may be their high affinity to translation initiation factors (Gallie, 1996).

It is possible that high translational activity of mRNA may be determined by some general contextual features (like to base composition or dinucleotide preferences). To detect the putative translational signals, we have analyzed the 5'UTRs of yeast mRNAs from the EMBL databank. Yeast mRNAs were choosen because they present a good model for investigation of translational signals: it is known that mRNAs of highly expressed genes is optimized at the level of translation elongation (by preferable using of synonymous codons corresponding to the major fractions of isoacceptor tRNAs; Sharp and Li, 1987). Here we describe general contextual features of yeast mRNA 5'UTRs and their correlation with CAI of mRNA coding parts.

Methods

Sequence data. The mRNA 5'UTR sequences were extracted from the EMBL database at the EMBL Internet site (http://www.embl-heidelberg.de/). Full-sized 5'UTRs were extracted from entries containing mapped transcription start sites and complete coding regions. Only mRNAs with experimentally determined transcription start sites were selected. If transcription was started from several sites, only shortest form of 5'UTR was extracted.

To avoid the bias due to redundant sequence data in the course of statistical analysis, sample of coding sequences were checked for redundancies. If two sequences in a sample share the similarity greater than 60%, then the shorter 5'UTR sequence is removed from the sample, while the longer one is retained for further

analysis. In total, 171 5'UTR sequences (100 5'UTRs started from single site and 71 5'UTRs presented the shortest form of leaders started from several sites). 171 sequences of 150 nt fragments upstream of transcription start sites (basal promoters) were used in comparative analysis to reveal 5'UTR-specific features.

Sequences analysis. System MGL (Kolpakov and Babenko, 1997) was used to analyze general contextual features of 5'UTR sequences (length, mono- and dinucleotide contents). Context of translation start codon was determined by using the simplest weight matrix ($W = |\ln w(b,j))|$, where w(b,j) is the weight of the nucleotide b in position j). The matching score $S(b_1,...,b_L)$ of a sequence $b_1,...,b_L$ was calculated as:

$$S(b_1,...,b_L) = \sum_{J=1}^{L} \ln w(b_j,j)$$

Codon adaptation index (Sharp and Li, 1987) was calculated by using a program CodonW (http://www.molbiol.ox.ac.uk/cu). Pearson's linear (LCC) and Kendall's rank tau (TAU) correlation coefficients were used for the measurement of correlation between two variables. These coefficients are based on different assumptions (parametric and nonparametric) and thus the relation between variables was analyzed independently.

Results

5'UTR of yeast mRNAs

Length of 5'UTR. Average length of yeast mRNA 5'UTR was found to be about 70±5.3 nt (only mRNAs with single transcription start site were used in this study). Analysis of distribution of 5'UTR lengths (data not presented) showed that the lengths of the most leaders (67%) vary between 20 and 80 nt.

Base composition. Nucleotide content in the yeast mRNA 5'UTRs is listed in the Table 1. For this analysis we used either 5'UTRs started from a single transcription start site (5'F) or extended sample (5'Fs) containing also the shortest leader if several ones had been annotated in the corresponding EMBL entry. Shortest leaders were selected because these sequences represent 5'UTRs *per se* whereas longer forms might possess features of both leader and basal promoter. One can see that nucleotide contents in samples 5'F and 5'Fs are very similar so we used more representative sample 5'Fs in the most kinds of analysis. Generally, 5'UTRs contain more A and less U than 5'untranscribed regions.

Base composition may affect secondary structure stability and, thereby, the movement of 40S ribosomal subunits along 5'UTR (Kozak, 1994). GC pairs have a major impact on the hairpin stability and therefore sequences containing more G and C have a potential to form more stable secondary structures. As can be seen from the Table 3, average G+C content in 5'UTRs is similar to that in promoters. However, if the sequence contains non-equal amounts of the complementary nucleotides, the possibility to form the stable secondary structure is lowered. G/C and A/U ratios (expressed as |G-C|/(G+C) and |A-U|/(A+U), respectively) were determined. It appeared that contents of the complementary nucleotides were considerably more asymmetric in the mRNA 5'UTRs than in promoters (coding strand). Another parameter reflecting opportunity to form stable secondary structure is E capacity (Ecap=9GC+3AU+2GU, where G,U,C,A – concentrations of nucleotides). Ecap of 5'UTRs was found to be lower then of promoters.

Table 1. Base composition (mean±standard error) in 5'untranslated and 5'untranscribed regions of yeast genes*.

	5'F	5'Fs	5'P
A	42.1±0.99	40.8±0.84	31.2±0.5
G	15±0.7	15.4±0.56	16.9±0.34
C	18.9±0.67	19.3±0.6	17.9±0.33
U	24.1±0.99	24.5±0.82	34±0.57
A+U	66.2±0.79	65.3±0.69	65.2±0.45
G/C	0.28±0.03	0.29±0.02	0.15±0.009
A/U	0.32±0.02	0.32±0.02	0.16±0.009
Ecap	0.59±0.01	0.60±0.01	0.69±0.005

*Parameters G/C, A/U and Ecap were calculated for the coding strand of 5'untranscribed region in the only aim compare its compositional organization with 5'UTR: these parameters cannot be used to describe DNA secondary structure of the promoter fragments.

It was found that the frequency of 5'UTRs with close contents of the complementary nucleotides ($0.75<G/C(A/U)<1.25$) was about 1.5-2 times lower for 5'UTRs compared to the promoter sense chains whereas portions of sequences with highly distinct content of complementary nucleotides (more than 2 fold) were considerably higher (Table 2). Thus, high asymmetry in the contents of G/C and A/U is a characteristic feature of mRNA leader regions. It is likely that this compositional characteristic was resulted from a selection against stable secondary structure that decreases 5'UTR functional activity and translation initiation rate.

Table 2. Ratio of the complementary nucleotides (distributions of *S. cerevisiae* mRNA 5'UTR sequences (%) with different A/U and G/C ratios)*.

Range	G/C		*A/U*	
	5'Fs	5'P	5'Fs	5'P*
<0.5,>2	33.9	6.9	43.3	8.7
0.75-1.25	36.3	53.8	25.7	43.9

*see footnote to the Table 1

Dinucleotide preferences. Dinucleotide contents in 5'UTRs and 5'untranscribed regions are shown in Table 5. One can see that 5'UTRs are characterized by considerable deviations in the contents of several dinucleotides from the expected values. In comparison with the coding strand of 5' untranscribed sequences 5'UTRs are characterized by deviations in dinucleotide usage: ApG and UpU are overrepresented (Obs/Exp>1.15), ApU and UpG are underrepresented (Obs/Exp<0.85). 5'UTRs are characterized by considerable asymmetry (>0.25) in the content of GA/AG, GC/CG, UG/GU.

Table 3. Average deviation of dinucleotide frequencies from the expected values expressed as observed to expected (Obs./Exp.) ratios*

	5'Fs	5'P		5'Fs	5'P
AA	1.06±0.02	1.14±0.02	CA	1.11±0.04	1.04±0.02
AG	1.31±0.05	1.01±0.03	CG	0.86±0.06	0.96±0.04
AC	0.97±0.03	0.90±0.02	CC	0.93±0.08	1.00±0.03
AT	0.80±0.02	0.93±0.02	CT	1.00±0.04	0.99±0.02
GA	0.98±0.04	1.05±0.03	TA	0.88±0.03	0.83±0.02
GG	0.88±0.07	1.08±0.04	TG	0.73±0.05	1±0.02
GC	1.13±0.07	1.12±0.04	TC	1.06±0.07	1.03±0.02
GT	0.98±0.05	0.86±0.02	TT	1.26±0.04	1.15±0.02

*For a dinucleotide a_1a_2, the Obs./Exp. ratio in a given sequence is calculated as $p(a_1a_2)/p(a_1)p(a_2)$, where $p(a_1)$, $p(a_2)$, and $p(a_1a_2)$ are the observed frequencies of the nucleotides a_1 and a_2 and the dinucleotide a_1a_2. Only sequences longer than 60 nt were considered.

Correlation between 5'UTRs contextual features and codon adaptation indices of coding sequences

To analyze 5'UTR features influencing translatability we determined their correlation with CAI. It was assumed that mRNAs of high expression genes possessing high rate of translation elongation (according to CAI) have to have 5'UTRs mediating an efficient interaction with translation initiation machinery. A correlation between the matching score of translational start codon context (from –6 to +3 nucleotides around AUG) and CAI is positive and highly significant (LCC = 0.42, TAU=0.31) that support the use of this approach. Correlation coefficients between CAI and 5'UTRs contextual features are listed in Table 4.

Table 4. Correlation between CAI and different features of 5'UTR*.

	5'UTRs (5'F, 95 sequences, >20 nt)							
	CAI^{LCC}	CAI^{TAU}		CAI^{LCC}	CAI^{TAU}		CAI^{LCC}	CAI^{TAU}
Length	-0.16	-0.04	*AA*	-0.04	0.02	*CA*	-0.10	-0.1
A	0.38	0.26***	*AG*	0.05	0.15*	*CG*	0.19	0.15*
G	-0.50	-0.34***	*AC*	0.08	0.04	*CC*	-0.05	-0.02
C	0.21	0.12	*AU*	0.09	-0.04	*CU*	0.07	0.14
U	-0.16	-0.08	*GA*	0.11	0.06	*TA*	0.02	0.07
G/C	0.38	0.11	*GG*	-0.38	-0.28***	*UG*	-0.07	-0.18**
A/U	0.21	0.14*	*GC*	-0.08	-0.11	*UC*	0.06	0.06
Ecap	-0.48	-0.30***	*GU*	0.03	0.09	*UU*	-0.11	-0.03

*Significant correlation (P<0,05) coefficients are bolded; For Kendall's rank TAU coefficient: ***P<0.001, **P<0.01; *P<0.05; dinucleotides are expressed as Obs./Exp. ratios.

As can be seen, several 5'UTR characteristics correlates with CAI: 5'UTRs of highly efficient mRNA contain more A and C and less G, they are highly asymmetric in the content of complementary nucleotides and characterized by lower Ecapacity values. As to preferences in dinucleotide usage, CAI is positively correlated with ApG and CpG and negatively with GpG (strongly) and UpG.

Discussion

The process of expression of eukaryotic genes consists from several consequence stages and results in a synthesis of a certain amount of functional protein. It may be supposed that high rate of expression can be reached if all expression stages occur at high efficiencies. Low efficiency of any expression process (DNA transcription, pre-mRNA processing, mRNA export, mRNA translation, or mRNA and polypeptide low cytoplasmic stabilities) is likely to be not compatible with high polypeptide production. Basing on this assumption we have supposed that the sequence characteristics of mRNAs of highly expressed eukaryotic genes are

optimized to support an efficient translation (Kochetov et al, 1998). According to this hypothesis, mRNA domains have to be optimized to provide high efficiencies of their specific functions: 5'UTRs features have to support a high rate of interaction with translation initiation machinery, coding parts have to provide high translation elongation rate. Functions of 3'UTRs in translation are not strictly determined, however, it may be supposed that these sequences participate in control of mRNA cytoplasmic stability. An analysis of correlations between some mRNA parameters supported this assumption: weight of start codon context correlates significantly with the codon adaptation index. This means, that an efficient initiation and elongation of translation are essential conditions of high expression rate. Similarly, contextual features influencing potential stability of mRNA secondary structure (asymmetry in the contents of the complementary nucleotides and E*capacity*) also correlates with CAI: namely, higher CAI correlates with lower potential to form stable secondary structures (Tables 4, 5).

Table 5. Features mRNA 5'UTRs potentially important to support high translation initiation rate.

5'UTR general features	Features correlating with mRNA translational activity
Mononucleotide content	*positively*: content of A and C, disbalance in the contents of the complementary nucleotides; *negatively:* content of G.
Dinucleotide preferences	*positively*: content of ApG, CpG; *negatively:* GpG, UpG

Note, that G content does correlates negatively with CAI whereas C content correlates positively: it may mean that G+C content (generally used to evaluate sequence potential to form secondary structure (e.g., see Kozak, 1994 and elsewhere)) may be incorrect criterion for single-stranded mRNA sequences). Interestingly, the length of yeast 5'UTRs is not correlates with CAI whereas these parameters were shown to be different between high and low expression mRNA of plants and mammals (Kochetov et al., 1998). It is likely, that this difference between yeasts and higher eukaryotes may result from the frequent occurrence of multiple transcription start sites in yeast promoters and multiple 5'UTR forms. The role of some 5'UTR features in translation process (e.g., dinucleotide preferences correlating with CAI significantly) is presently unclear: however, these data provide a good background for the next experimental analysis.

Acknowledgements

A. Kochetov was supported by SB RAS grant for young scientists and in the frames of RFBR grant for support of scientific school of V.K. Shumny.

References

1. Gallie,D.R (1996) Translational control of cellular and viral mRNAs. *Plant Mol. Biol.* **32**, 145-158.
2. Futterer,J. and Hohn,T. (1996) Translation in plants - rules and exceptions. *Plant Mol. Biol.* **32**, 159-189.
3. Kochetov, A.V., Ischenko, I.V., Vorobiev, D.G., Kel, A.E., Babenko, V.N., Kisselev, L.L., Kolchanov, N.A. (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett* 440: 351-355.
4. Kolpakov, F.A., Babenko, V.N. (1997) Computer system MGL: tool for sample generation, visualization and analysis of regulatory genomic sequences. Mol Biol (Moscow) **31**: 647-655.
5. Kozak,M. (1994) Determinants of translational fidelity and efficiency in vertebrate mRNAs. *Biochimie* **76**, 815-821.
6. Pain,V. (1996) Initiation of protein synthesis in eukaryotic cells. *Eur. J. Biochem.* **236**,747-771.
7. Sharp, P. M., and Li, W. H. (1987). The codon adaptation index a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* **15**, 1281-1295.

COMPOSITIONAL PROPERTIES OF PLANT mRNA 5'UNTRANSLATED REGIONS: THE PRESENCE OF ENHANCER-LIKE MOTIFS

Kochetov A.V., Glazko G.V., Sirnik O.A., Rogozin I.B., Trifonova E.A., Komarova M.L., Shumny V.K.

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

Keywords: mRNA, sequence characteristics, 5'UTRs, translation, enhancers

Resume

Motivation*:*

It is well known that the efficiency of eukaryotic mRNA translation initiation depends on sequence characteristics of the 5'untranslated region (5'UTR). However, 5'UTR-located signals influencing initiation efficiency are still poorly investigated. Here we present the results of comparative analysis of 5'UTRs, introns, 3'UTRs, and 5'untranscribed regions by the content of simple oligotracts characteristic for some translational enhancers.

Results*:*

Generally, 5'UTRs of dicot mRNAs are characterized by the strong deviation of dinucleotide content from the expected ones and differs from the other functional sequences. Despite neither simple oligotract was found to present at very high frequency, several oligotracts are considerably depleted. These data allow hypothesizing that sequence organization of 5'UTRs is not random and may be adapted to specific demands of translation initiation machinery.

Introduction

It is known that translational efficiency of eukaryotic mRNAs varies considerably with their sequence characteristics (for review, see Kozak, 1994; Futterer and Hohn, 1996, Pain, 1996). Contextual and structural features of the mRNA 5' untranslated region (5'UTR) significantly affect the rate of translation initiation and, thereby, the level of polypeptide production. It is widely accepted that the majority of eukaryotic mRNAs are translated through the linear scanning mechanism (for review, see Kozak, 1994). According to this model, some features of the leader sequence influence mRNA translational efficiency, i.e., the context of translational start codon, the presence of AUGs within 5'UTR, and the stable secondary structure in the leader. Apart from the context of the start codon, influencing its recognition efficiency, eukaryotic mRNA leader sequences are considered to be highly divergent: none conservative translational signals were found (Pain, 1996). Nevertheless, there exist a rapidly growing number of experimental evidences showing the active role of 5'UTR in translation initiation process. For example, stress-specific translation of *Drosophila* and plant HSP mRNAs depend on unknown signals located within their 5'UTRs; tissue-specific increase of translational efficiency of the barley alpha-amylase mRNA in endosperm and aleurone layer cells depends on its untranslated regions. Translational enhancers are one of the most interesting phenomena of this sort. It was experimentally found that 5'UTRs of several plant viruses (tobacco mosaic virus, alfalfa mosaic virus RNA 4, brome mosaic virus RNA 4, potato X virus, 5'UTRs of tobacco etch virus; and some others) are able to enhance translational efficiency of the downstream coding sequences. Mechanisms of their specific activity are not clear yet, however, one of the possible reasons may be their high affinity to translation initiation factors (Gallie, 1996).

Thus, a phenomenon of translational enhancers clearly demonstrates that nucleotide sequence of 5'UTR may influence the efficiency of their interaction with cellular translational machinery and determine mRNA translational rate. However, known translational enhancers contain no common element that may serve as a general translational signal (like to Shine-Dalgarno site in prokaryotic mRNAs). We suppose that nucleotide composition of mRNA 5'UTR may influence its translational activity (presumably due to varying affinity to translation initiation factors) but its active element(s) may be represented by some kind of oligotracts or repeats of variable structure. For example, the results of mutagenesis experiments with tobacco mosaic virus translational enhancer revealed two active sites within 68-nt long sequence: three 8-nt repeats *acaattac* and *(CAA)*-tract (Gallie and Walbot, 1992). Similarly, it has been proposed that *CCACC*-motif may also be an active element of potato X-virus 5'UTR (Tomashevskaya et al., 1993); alfalfa mosaic virus enhancer contains several oligo*(U)* stretches, etc. It may be possible that various nucleotide compositions within 5'UTR provide mRNA with high translational activity: this fact could explain both non-similarity of translational enhancers and high extent of

plant leader variability. To detect the putative translational signals, we have analyzed the 5'UTRs of plant mRNAs from the EMBL databank by measuring the content of simple oligorepeats.

Methods

Sequence data. Full-sized 5'UTRs (5'F) were extracted from the entries containing sequences of genomic DNA clones with mapped transcription start sites. A sample of possibly incomplete 5'UTRs (5'I) was also created using the EMBL entries describing cDNA sequences (only sequences exceeding 15 nt were taken into analysis) (Table 1). To avoid the bias due to redundant sequence data in the course of statistical analysis, both 5'F and 5'I samples were checked for redundancies. If two sequences in a sample share the similarity greater than 70%, then the shorter sequence is removed from the sample, while the longer one is retained for further analysis. Besides, several data sets of nucleotide sequences from the other gene regions (Table 1) involved in different expression processes were created. Samples of mRNA 3'UTRs (3'U, sequences stretching from the translational stop codon to the site of polyadenylation), proximal (i.e., closest to the 5' end) introns (In) and 5' untranscribed regions (5'P, 250 nt upstream transcription start site) of plant genes were also created (Table 1). All these sequences were extracted from the plant genes that were used as a source for the experimentally mapped complete 5'UTRs (sample F). Only dicot 5'UTR mRNA sequences were taken for analysis, because of their higher representation in comparison with monocots.

Table 1. Description of the 5'UTR samples used in this study

	5'UTRs(5'F)	5'UTRs(5'I)*	3'UTRs(3'U)	5' introns (In)	Promoters (P)
Dicots	184	2475	115	147	160

*Length > 15 nt

Sequences analysis. The frequencies of different motifs were calculated on the basis of the *ad hoc* programs. The motifs with frequencies deviating from the standard normal expectation at the 5% level of significance were considered as "overrepresented", that is, $f_{obs}>\mu+1{,}96*\sigma$. This criterion was used here only as the measure of deviation, but not in order to reveal the frequencies deviating from the normal distribution. The motif's density per 100 nt was calculated as $\rho=100*\Sigma(n_{iA})/\Sigma(\lambda_i)$, where n_{iA} is the number of 'A' motifs in the i-th sequence, λ_i is the length of the i-th sequence.

Results and discussion

Functional domains of plant gene are characterized by distinct nucleotide content and dinucleotide preferences (Table 2). Compared to 5'UTRs, 3'UTRs and introns contain considerably more uridines and less cytidines, whereas their nucleotide contents in the leader sequences and 5' untranscribed regions are similar.

Compared with the other sequences, dinucleotides ApT and CpC are specifically underrepresented (the ratio of observed to expected frequencies is <0.85) in the 5'UTRs. Overrepresented dinucleotides (>1.15) are ApG, TpC, and TpT. Interestingly, dinucleotide preferences along 5'UTRs was different: 5'- and 3'-terminal 21 nt-fragments of 5'UTRs strongly differ in the Obs/Exp frequencies of CpA, CpC, TpG, TpT; CpA, and ApC.

Table 2. Average nucleotide and dinucleotide content (%) of dicot mRNAs 5'-untranslated sequences in comparison with the sequences with other functions (dinucleotide contents are expressed as observed to expected ratios*).

Set	A	G	C	U	AA	AG	AC	AT	GA	GG	GC	GT	CA	CG	CC	CT	TA	TG	TC	TT
5'F	35	13	23	29	1.15	1.42	0.86	0.77	1.09	0.89	0.99	0.95	1.19	0.56	0.69	1.19	0.66	0.92	1.37	1.2
5'I	32	16	23	29	1.23	1.3	0.81	0.78	1.27	0.96	0.92	0.86	1.12	0.77	0.81	1.15	0.59	0.86	1.39	1.26
$5'^1$	36	10	26	28	1.04	1.19	0.90	0.80	1.10	0.93	0.80	0.91	1.17	0.55	0.72	1.04	0.59	1.03	1.29	1.09
$5'^2$	39	15	20	26	1.13	1.22	0.74	0.72	1.17	0.92	0.77	0.82	1.00	0.71	0.95	1.05	0.55	0.80	1.35	1.28
3'U	30	17	14	39	1.14	0.89	0.9	0.91	0.94	0.88	1.05	1.07	1.09	0.52	1	1.12	0.81	1.29	1.03	0.99
In	28	16	14	42	1.13	0.94	0.98	0.95	1.1	0.77	0.9	0.99	1.18	0.69	0.93	1.04	0.81	1.15	1.07	1.03
5'P	34	14	20	32	1.12	1	0.92	0.9	0.98	1.11	0.92	1.04	1.07	0.68	1.14	0.97	0.82	1.12	1.04	1.1

* Only sequences with the length at least 40 nt were considered; 21 nt fragments starting from 5'end of mRNAs (sample 5'F) and 21 nt fragments located upstream the start AUG codons (sample I) were concatenated and accumulated in the samples $5'^1$ and $5'^2$, respectively.

Table 3. Comparative analysis of simple oligotract occurrence within plant gene sequences.

	5'UTR 203	3'UTR 212	5'introns 271	promoters 280
(a)	E(a)=0.13±0.34	E(a)=0.07±0.25	E(a)=0.10±0.3	E(a)=0.18±0.38
(g)	E(g)=0	E(g)=0	E(g)=0	E(g)=0.01±0.08
(t)	E(t)=0.04±0.21	E(t)=0.12±0.33	E(t)=0.25±0.43	E(t)=0.09±0.29
(c)	E(c)=0.01±0.1	E(c)=0	E(c)=0	E(c)=0
(tc)	E(tc)=0.67±0.47	E(tc)=0.61±0.49	E(tc)=0.66±0.48	E(tc)=0.64±0.48
(ag)	E(ag)=0.43±0.50	E(ag)=0.38±0.49	E(ag)=0.35±0.48	E(ag)=0.55±0.50
(tg)	E(tg)=0.20±0.40	E(tg)=0.69±0.47	E(tg)=0.76±0.43	E(tg)=0.43±0.50
(at)	E(at)=0.50±0.50	E(at)=0.93±0.26	E(at)=0.87±0.34	E(at)=0.95±0.22
(gc)	E(gc)=0.02±0.16	E(gc)=0	E(gc)=0.01±0.09	E(gc)=0.03±0.17
(ac)	E(ac)=0.50±0.50	E(ac)=0.27±0.44	E(ac)=0.34±0.47	E(ac)=0.67±0.47

Content of oligotracts. Plant sequences were analyzed by simple mono- and dinucleotide motifs (type (n1,n2), where the ratio n1/n2 may vary: (a), (g), (c), (t), (t,c), (a,g), (t,g), (a,t), (g,c), (a,c); the length of motifs is at least 8 nt). It was found that (i) the 5'UTRs differ considerably from 3'UTRs, 5'untranscribed region and 5'proximal introns by the frequency of (t), (t,g), (a,t), (underrepresented), (ii) (a,c) is overrepresented in 5'UTRs and 5'untranscribed regions (Table 3). Notably, the density of oligotracts (t,c) and (a,g) in 5'UTRs is considerably higher than in the other sequences, whereas the density of (c,a) is higher for both 5'UTRs and 5'untranscribed regions (data not shown).

Summary

The results of plant genes analysis by simple oligotract content have demonstrated: **(1)** 5'UTRs differ from the other gene domains by content of several dinucleotides. Such function-dependent difference may be explained by the influence of these dinucleotides on the 5'UTR translational properties. Generally, observed dinucleotide frequencies in 5'UTRs were found more deviating from the expected ones than the other sequences. **(2)** 5'UTRs do not contain obligatory oligotract (n1,n2) (whereas the other functional sequences do contain (a,t) and (t,g) oligotracts very frequently). Oligotract (a,c), which is an active element of the TMV translational enhancer, is more frequently present in 5'UTRs than in introns or 3'UTRs. However, 5'UTRs were found to contain several oligotracts (e.g., (t), (t,g), (a,t)) in considerably lower concentrations in comparison with the other sequences. **(3)** It may be suggested that different domains of 5'UTR perform different functions: 5'terminal domain is involved into interaction with translation initiation factors and loading of 40S ribosomal subunits, 3'terminal domain (AUG codon context) is responsible for recognition of start codon by a ribosome, and the middle part of the 5'UTR interacts with the scanning 40S ribosomal subunits. It seems possible that compositional features of these domains may differ because they support different functions. Analysis of 5' and 3' leader ends proves this assumption – these domains are clearly differ by occurrence of CpA, CpC, UpG, and UpU. Frequencies of simple oligotracts within 5' and 3'terminal domains of 5'UTRs were also calculated (Table 4). It was found that the contents of some oligotracts vary: (a,c) is more frequent near the cap-site, whereas (a,g) – near the start codon.

Table 4. Frequency of the simple oligotracts (longer than 5 or 8 nt) within 5' and 3'-terminal domains of the leader sequences.

	>8, 21 nt prior AUG	>5, 21 nt prior AUG	>8, 21 nt after cap-site	>5, 21 nt after cap-site
A	E(a)=0.04±0.21	E(a)=0.14±0.34	E(a)=0.02±0.14	E(a)=0.08±0.28
C	E(c)=0	E(c)=0	E(c)=0	E(c)=0.01±0.12
G	E(g)=0	E(g)=0	E(g)=0	E(g)=0
T	E(t)=0	E(t)=0.04±0.19	E(t)=0	E(t)=0.02±0.14
(TC)	E(tc)=0.10±0.30	E(tc)=0.32±0.47	E(tc)=0.15±0.36	E(tc)=0.33±0.47
(AG)	**E(ag)=0.13±0.33**	**E(ag)=0.34±0.48**	**E(ag)=0.06±0.24**	**E(ag)=0.19±0.39**
(TG)	E(tg)=0.03±0.18	E(tg)=0.15±0.36	E(tg)=0.02±0.16	E(tg)=0.13±0.33
(AT)	E(at)=0.12±0.32	E(at)=0.37±0.48	E(at)=0.09±0.29	E(at)=0.31±0.46
(GC)	E(gc)=0.01±0.12	E(gc)=0.03±0.17	E(gc)=0	E(gc)=0.03±0.17
(AC)	**E(ac)=0.09±0.28**	**E(ac)=0.27±0.44**	**E(ac)=0.16±0.36**	**E(ac)=0.44±0.50**

Acknowledgements

A. Kochetov was supported by SB RAS grant for the young scientists. V.K. Shumny has received a RFBR grant for the support of scientific schools.

References

1. Gallie,D.R (1996) Translational control of cellular and viral mRNAs. *Plant Mol. Biol.* 32, 145-158.
2. Gallie,D.R. and Walbot,V. (1992) Identification of the motifs within the tobacco mosaic virus 5'-leader responsible for enhancing translation. *Nucleic Acids Res.* 20, 4631-4638.
3. Futterer,J. and Hohn,T. (1996) Translation in plants - rules and exceptions. *Plant Mol. Biol.* 32, 159-189.
4. Kozak,M. (1994) Determinants of translational fidelity and efficiency in vertebrate mRNAs. *Biochimie* 76, 815-821.
5. Pain,V. (1996) Initiation of protein synthesis in eukaryotic cells. *Eur. J. Biochem.* 236,747-771.
6. Tomashevskaya O. L., Solovyev A. G., Karpova O. V., Fedorkin O. N., Rodionova N.P., Morozov S.Yu., Atabekov J. G. (1993) Effects of sequence elements in the potato virus X RNA 5'non-translated alpha beta-leader on its translation enhancing activity *J. Gen. Virol*. 74, 2717-2724.

TRANSLATIONAL FEATURES OF 5'UTR-LOCATED MINIORFS

[1]Kochetov A.V., [1,2]Sirnik O.A., [1]Komarova M.L., [1]Trifonova E.A., [1]Kolchanov N.A., [1]Shumny V.K.

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

Keywords: mRNA, 5'UTR, miniORF, main ORF, translational features

Resume

Motivation:

It is known that 5'UTR-located AUGs and small ORFs beginning form these AUGs (miniORFs) can decrease the level of translation of the main ORF. However, miniORF features influencing mRNA translatability are poorly studied. The aim of the work presented is to determine miniORF characteristics potentially important for mRNA translation in dicot plant cells.

Results:

Within the sample of 5'UTR of dicot mRNAs extracted from the EMBL databank, the upstream miniORFs were analyzed. Conclusions as follows were made. (1) 5'UTRs containing upstream miniORFs comprise unexpectedly large portion of the sample (about 20%). (2) Most upstream AUGs are located within non-optimal context and encode small ORFs, which are likely to decrease their negative influence on mRNA translatability. (3) It was found that there exists a relationships between translational signals of miniORFs and those of the main coding sequences: more optimal context of the main coding sequence start codon corresponds to the less optimal context of miniORF AUG codon and more optimal context of miniORF stop codon. It is likely that decrease of miniORF translation level can increase the main ORF translation rate (presumably due to the "leaky scanning"), whereas an efficient termination at miniORF stop codon is essential for high-rate traffic of 40S ribosomal subunits through mRNA 5'UTR.

Introduction

It is known that various eukaryotic mRNAs possess by different translation rate depending on their sequence characteristics (Futterer and Hohn, 1996). mRNA 5'untranslated region (5'UTR) participates in interaction with eIFs and 40S ribosomal subunits and determines an efficiency of translation initiation (Ray et al., 1983; Kozak, 1994; Futterer and Hohn, 1996; Kochetov et al., 1998). It is widely accepted that most eukaryotic mRNAs are translated by a linear scanning mechanism (Kozak, 1994). In the frames of this model, 40S ribosomal subunit interacts with the cap at mRNA 5'terminus and moves along 5'UTR until first AUG triplet, whereupon 60S subunit binds and translation begins. If mRNA contains 5'proximal AUG(s) upstream the main ORF, some 40S ribosomal subunits can recognize them and translate corresponding miniORFs (as a rule, very small, miniORFs). If such an upstream AUG falls within an optimal context (i.e., adenine occurs in position –3 (Kozak, 1994; Futterer and Hohn, 1996), it will be recognized by the most ribosomes loaded onto this mRNA. In this case, mRNA could be translated (generally at lower efficiency) by reinitiation mechanism (Kozak, 1994). If the context of upstream AUG is unoptimal, some 40S ribosomal subunits will not recognize it; they continue scanning in 3'direction and can initiate translation at the main ORF (Kozak, 1994). According to the experimental data, the presence of miniORF within a 5'UTR of functional mRNA decreases its translational efficiency (except of some rare cases of specific initiation mechanisms, i.e., internal initiation or ribosomal shunt (Futterer and Hohn, 1992; Kozak, 1994; Futterer and Hohn, 1996)). Negative effect of miniORF depends on its features (AUG context, length, and position relatively the main ORF) (Futterer and Hohn, 1992). Nevertheless, mechanisms of influence of upstream AUGs and miniORF features on translation initiation are not clear yet (especially for plant mRNAs, since the most experiments were performed on animal translational system (Kozak, 1994; Futterer and Hohn, 1996)). The aim of this work is to analyze mRNA 5'UTRs of dicot plant mRNAs (extracted from the EMBL databank) for the presence of miniORFs and to determine their characteristics potentially influencing mRNA translatability.

Methods

1. Full-sized 5'UTRs (5'F) were extracted from the EMBL entries containing sequences of genomic DNA clones with mapped transcription start sites. A sample of potentually incomplete 5'UTRs (5'I) was also created by using the EMBL entries, which describe cDNA sequences (only the sequences exceeding

15 nt were taken into analyses) (Table 1). To avoid the bias due to redundant sequence data in the course of statistical analysis, both 5'F and 5'I samples were checked for redundancies. If two sequences in a sample share the similarity greater than 70%, then the shorter sequence is removed from the sample, while the longer one is retained for further analysis. Besides, several data sets of nucleotide sequences from the other gene regions (Table 1) involved in different expression processes were created. Samples of mRNA 3'UTRs (3'U, or the sequences stretching from the translational stop codon to the site of polyadenylation) and proximal (i.e., the closest to the mRNA 5' end) introns (In) of plant genes were also created (Table 1). All these sequences were extracted from the plant genes that were used as a source for the experimentally mapped complete 5'UTRs (sample F). Only dicot 5'UTR mRNA sequences were taken for analysis, because of their higher representation in comparison with monocots.

Table 1. Samples description.

Name of a sample	5'UTR (I)	5'UTR (F)	3'UTR	Introns (I)
Number of sequences	2475	333	69	271

Sample of miniORF sequences (created by using MGL package (Kolpakov and Babenko, 1997)) was extracted from the 5'UTRs containing 1 or 2 upstream AUGs. We have supposed that higher upstream AUG content might be hardly compatible with an efficient translation by the linear scanning mechanism, so that such sequences could be non-functional or translated by the distinct mechanisms. The sample of miniORFs was compiled of 333 sequences extracted from both full-sized and incomplete dicot mRNA 5'UTRs.

2. System MGL was used to analyze general contextual features of 5'UTR sequences (length, mono- and dinucleotide contents). Context of translation start codon was determined by using the simplest weight matrix ($W = |\ln w(b,j))|$, where w(b,j) is the weight of the nucleotide b in position j). The matching score $S(b_1,...,b_L)$ of a sequence $b_1,...,b_L$ was calculated as:

$$S(b_1,...,b_L) = \sum_{J=1}^{L} \ln w(b_j,j).$$

Context positions taken into account for the start codon were –3 and +4 (major determinants of AUG initiation activity) and those for the stop codon were spanning from –3 to +3 (including the second and third positions of the termination codon itself). Consensus weight matrices for initiation and termination signals were used according to TransTerm database data on *Arabidopsis thaliana* (http://biohem.otago.ac.nz:800/Transterm). Pearson's linear (LCC) and Kendall's rank tau (TAU) correlation coefficients were used for the measurement of correlation between two variables. These coefficients are based on different assumptions (parametric and nonparametric) and, therefore, the relation between the variables was analyzed independently.

Results and discussion

Upstream miniORFs of dicot mRNAs and their characteristics upstream AUG frequency. It was found that *ca*. 20% of dicot mRNA 5'UTRs contain one (9-10%), two (4-6%), or more AUG codons (Table 2), whereas 3'UTRs and proximal introns contain more AUGs with considerably higher frequencies. Comparison of observed AUG frequencies with the expected values (i.e., in the case of the random codon choice without selection) has indicated that the content of upstream AUGs in 5'UTRs is considerably depleted. However, the observed upstream AUG frequency appears to be considerably more than it was supposed earlier (about 10% according to (Kozak, 1994; Futterer and Hohn, 1996).

Table 2. An occurrence of upstream AUGs within various dicot gene domains.

Sample		5'UTR (I)	5'UTR (F)	Introns	3'UTR
AUG-containing sequences (%)		19	20	92	95
Portions of sequences (%) containing	1	10	9	11	6
	2	4	6	11	6
	3	2	2	10	15
	4	1	2	8	13
	>4 AUGs	2	1	52	54
Obs/Exp* ratio		0.2	0.3	1.1	1.3

Obs – observed number of upstream AUGs; Exp –number expected due to the random codon choice without selection (Exp=A*U*G*(L-3), where A,U,G – frequency of nucleotides, L – sequence length, nt)

Length of miniORFs. Most miniORFs (about 70%) are less than 51 nt in length, whereas the average length is 42 nt.

Upstream AUG context. It was found that upstream AUGs (in contrast to the start AUGs of the main ORFs) usually lie in unoptimal context (Table 3). Adenines at position –3 are present in 29-33% of miniORFs, guanines at +4 position – in 23-26% of miniORFs, this percentage being several-fold less than in the context of the start AUG codon of the main ORF.

Interdependence of translational features between mini- and main ORFs. Relationships between some translational features of ORFs (weights of initiation and termination codon contexts) were analyzed. These parameters were chosen for analysis, because they represent the already known translational signals influencing mRNA translational activity in model experiments. Weight of start codon reflects the extent of similarity between certain site and consensus: it was shown that consensus of AUG codon context corresponds to its initiator activity (Futterer and Hohn, 1996; Joshi et al, 1997; McCaughan et al, 1995). The results are illustrated in Table 4:

Table 3. Nucleotide content in positions –3 and +4 of contexts of start codon of the upstream miniORFs and the main ORFs.

Context position	-3				+4			
nucl. content (%)	A	G	C	U	A	G	C	U
miniORFs (sample F)	33	17	17	30	34	23	21	21
miniORFs (sample I)	29	18	17	34	31	26	18	24
main ORFs (sample F)	67	18	5	11	15	70	4	11
main ORFs (sample I)	57	24	8	12	16	67	5	12

Table 4. Correlations between translational signals of miniORFs and mainORFs of dicot plant mRNAs.

Correlations between the weights of:	TAU	LCC
start codon contexts of miniORFs and main ORFs	**-0,12** (P=0,003)*	0,04
stop codon context of miniORFs and start codon context of the main ORFs	**0,15** (P=0,0002)*	**0,17**

* significance level; significant correlation coefficients are given in bold.

1) correlation between the weights of start AUGs of miniORFs and main ORFs is significant and negative. It may mean that high translational activity is supported by an efficient initiation at the start codon of the main ORF and inefficient initiation at the start codon of upstream miniORF. In this situation (according to the scanning model), a portion of 40S ribosomal subunits will pass upstream AUG ("leaky scanning") and continue scanning until the main ORF will be reached;

2) correlation between the weights of miniORF stop codon and main ORF start codon is significant and positive. This fact is unexpected, because the role of termination codon context in plant cells is poorly known. It may be assumed that ribosomes delay at unoptimal termination signal that stops the ribosome flow and decreases the scanning rate;

3) It may be assumed that the weights of start and stop codons of miniORFs can be interrelated too. It was found that these parameters correlate negatively (as can be expected from the correlations described above), but insignificantly (TAU: -0,05; significance level: 0,19). Nevertheless, these data allow us to suppose how natural selection decreases the negative influence of upstream AUGs and miniORFs encoded. Start codon of miniORFs tends to be less efficient, whereas stop codon tends to be highly efficient. It can increase the number of ribosomes reaching the genuine translational start site by leaky scanning and decrease the time of ribosome stoppage at miniORF stop codon.

Conclusion. It was found that 5'UTR-located miniORFs of dicot mRNAs possess by non-random characteristics, which may mediate their effect on mRNA translational activity. It is likely that negative influence of miniORFs is lower, if its start codon lies in non-optimal context (according to the scanning model by Kozak, an efficiency of main ORF translation may be increased by a leaky scanning mechanism). Unexpectedly, the weight of miniORF stop codon correlates significantly and positively with the weight of AUG codon of the main ORF. This fact allows assuming that this parameter also plays an essential role in determination of mRNA translation rate. First, it may be suggested that the "strength" of the stop codon can influence the time of delay of a ribosome translated from miniORF at its stop codon. Such ribosome will stop the movement of 40S ribosomal subunits along mRNA and, thereby, decrease translation rate of the main ORF. Second, the stoppage of 40S ribosomal subunits upstream miniORF (resulted from an inefficient termination at the stop codon of miniORF) may be followed by more efficient initiation at the AUG codon of miniORF (e.g., see a review by Kozak, 1994). This can increase the miniORF negative potential and cause a more pronounced decrease in polypeptide synthesis rate.

Acknowledgements

A. Kochetov was supported by SB RAS grant for young scientists; this work was partially sponsored within the frames of RFBR grant supporting scientific school of V.K. Shumny. The authors are grateful to G. Orlova for the help in translation of the manuscript.

References

1. Futterer, J., Hohn, T. (1992) Role of an upstream open reading frame in the translation of polycistronic mRNAs in plant cell. Nucl. Acids Res. **20**, 3851-3857.
2. Futterer, J. and Hohn, T. (1996) Translation in plants - rules and exceptions. *Plant Mol. Biol.* **32**, 159-189.
3. Joshi, C.P., Zhou, H., Huang, X., Chiang, V.L. V (1997) Context sequences of translation initiation codon in plants. Plant.Mol. Biol.**35**, 993-1001.

4. Kolpakov, F.A., Babenko, V.N. (1997) Computer system MGL: tool for sample generation, visualization and analysis of regulatory genomic sequences. Mol Biol (Moscow) **31,** 647-655.
5. Kochetov, A.V., Ischenko, I.V., Vorobiev, D.G., Kel, A.E., Babenko, V.N., Kisselev, L.L., Kolchanov, N.A. (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS Lett* **440,** 351-355.
6. Kozak,M. (1994) Determinants of translational fidelity and efficiency in vertebrate mRNAs. *Biochimie* **76,** 815-821.
7. McCaughan, K.K., Brown, C.M., Dalphin, M.E., Berry, M.J., Tate, W.P. (1995) Translation termination efficiency in mammals is influenced by the base following the stop codon. Proc. Natl. Acad. Sci. USA. **92**, 5431-5435.
8. Ray,B.K., Brandler, T.G., Adya, S., Daniels-McQeen, S., Miller, J.K., Hershey, J.W.B., Grifo, J.A., Merrick, W.C., Thach, R.E. (1983) Role of mRNA competition in regulation translation: Further characterization of mRNA discriminatory initiation factors. Proc. Natl. Acad. Sci. USA. 80, 663-667.

ASSOCIATED WITH REPEATED ELEMENTS STRUCTURAL DEFORMATION OF PROMOTER DNA UPON TRANSCRIPTION COMPLEX FORMATION

Masulis I.S., Chasov V.V., Ozoline O.N.

Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: ozoline@venus.iteb.serpukhov.su
*Corresponding author

Keywords: transcription complex formation, promoter, DNA structure

Resume

Motivation:

Adaptive conformational transitions of DNA accompanying protein-DNA interaction are now considered as a structural basis underlining stability and functional properties of protein-DNA complexes (Jones et al., 1999). The nature and direction of ligand-induced perturbations are potentially determined by sequence-dependent fine structural peculiarities of the free DNA, thus increasing the significance of DNA structural modeling. However the detailed studies are still required to estimate the extent of these correlations for particular DNA-protein complexes. The present study is targeted to experimental detection of DNA structural rearrangements, which take place upon transcription complex formation by *E.coli* RNA polymerase with promoter D of bacteriophage T7. The nucleotide sequence of this promoter possesses a set of periodically distributed direct repeats with high propensity to form unusual tertiary structure. To reveal promoter regions with altered DNA conformation both in the free DNA and transcription complexes, $KMnO_4$ footprinting technique was applied.

Results:

A pronounced helix deformation was detected in the structure of T7D DNA complexed with RNA polymerase. This distortion is located distantly from the transcription start point and takes place in the region of repeating element, which exhibits sensitivity to KMnO4 in the free promoter DNA. Overall conformation of the transcription complexes could therefore be controlled by the sequence-dependent structure of free promoter DNA.

Introduction

An individual conformational characteristics of the complexes formed by RNA polymerase with distinct promoters are highly informative in terms of sequence-dependent realization of the general principles in the organization of transcription machinery. On the other hand, a detailed analysis of transcription complexes may have a heuristic significance allowing revealing basically new features. In spite of the apparent uniformity in the topology of the complexes formed by the enzyme at different promoters (the data of a comparative computer analysis, Ozoline and Tsyganov, 1995), specific structural variations may be identified for any given promoter. Structural-functional relations underlying the complicated process of transcription initiation could be deduced on the basis of extensive studies of RNA polymerase-promoter complexes by different methods of fine structural probing. In some cases the data obtained allow extracting the sequence modules of promoter DNA determining structural “image” of transcription complexes and lighten the intrinsic mechanisms of promoter functioning.

One of the early promoters of bacteriophage T7 – D – was selected for this analysis. One of remarkable features of its primary structure is a presence of periodically distributed repeated elements with a context CTTTAGG (Fig.1). Structural mapping of free promoter DNA revealed a set of S1-nuclease-sensitive sites. Their positioning coincides with the location of direct repeats along this template (Masulis et al., 1998). In the present study we tried to follow, how do the regions with altered DNA conformation behave upon interaction with RNA polymerase. KMnO4 was used as a reagent, specifically modifying thymidines in the non-paired or unstacked state. Standard DNase1 footprinting assay was performed as a control to confirm complex formation and to localize the borders of promoter surface, interacting with the enzyme.

Methods

355 bp DNA fragment containing T7D promoter was prepared by PCR amplification. α32P – dCTP was incorporated in position –117 of the template strand after restricton of the promoter DNA by Hin1I and subsequent filling of the sticky end.

RNA polymerase was reconstituted from isolated subunits.

RNA polymerase-promoter complexes, containing ~10-6 M of enzyme and 0.2×10-6 M promoter DNA were preformed at in transcription buffer: 0.01M Tris-HCl, pH 7.9; 10 mM MgCl2; 25 mM KCl.

KMnO4- treatment of transcription complexes was performed according to Zaychikov et al., 1997; DNAse1 footprinting assay was performed according to Masulis et al., 1998.

Results and discussion

KMnO4 –footprinting technique is a commonly used approach to map unpaired bases in close vicinity to the transcription start site. Additional sites of reactivity, indicating local structural deformations, induced by RNA polymerase, were also revealed. For example, they were observed at the positions -56 and +20 of galP1 (Burns et al., 1996), -56 of cysG (Belyaeva et al., 1993) and +18 of rrnbP1 (Newlands et al., 1991). These structural alterations are located near the boundaries of promoter contact surface and usually are considered as a result of direct mechanical distortions of promoter DNA by the enzyme molecule.

In the case of T7D an expected deformation reflecting the formation of the transcriptionally competent open complex was registered near the transcription start point (-11~-4) (Fig.2). At the same time, a pronounced distortion was found in the coding part of the gene (+43~+51). This additional region exhibiting reactivity to KMnO4 lies two helical turns downstream from the contact area with RNA polymerase (see DNAse1 footprinting ladder) which is not typical for any other transcription complexes. The position of this non-contact structural deformation corresponds to the last of 12 periodically repeated elements with the sequence context CTTTAGG (Fig. 1). It was previously suggested that repeated motif directly participates in the formation of three-dimensional structural "image" of T7D promoter, significant for its recognition by RNA polymerase (Masulis et al., 1998). Periodical pattern of different nature was found in many other promoters and thus is not an exceptional property of T7D. In the free DNA neighboring repeated elements may adopt a special conformation known as "slipped-loop structure" (SLS) (Khomyakova et al., 1998; Pearson et al., 1998) that is due to alternative complementary interaction of opposite strands. Single-strand-specific chemical and enzymatic probing could identify these elements. The presence of unpaired bases correlating in location with distribution of repeated segments in the structure of T7D was really detected by S1 nuclease (Masulis et al., 1998) or KMnO4 probing (Fig.2). Observed distribution of structural abnormalities is highly reproducible and therefore reflects steady state of this DNA template. The fact that conformational probing does not reveal one preferred location of helix distortion suggests a possibility that this DNA adopts a variety of structural isomers. That is in line with a possibility of the formation of SLS isomer by different repeated modules resuming balanced equilibrium at given experimental conditions.

Obvious increase in reactivity to $KMnO_4$ has been detected at the region +43~+51 when the promoter DNA was subjected for interaction with RNA polymerase. Dynamic structural rearrangement, which takes place upon transcription complex formation results in redistribution of structurally distorted sites and one preexisted conformation of promoter DNA with unpaired bases in +43~+51 is selected as dominant. Thus, repeat-associated tertiary constructions may undergo cooperative rearrangement and affect the final configuration of transcription complex. Therefore, the presence of direct repeats in promoter DNA and early-transcribed region should be taken into account in assessment of functional properties of natural and artificial promoters.

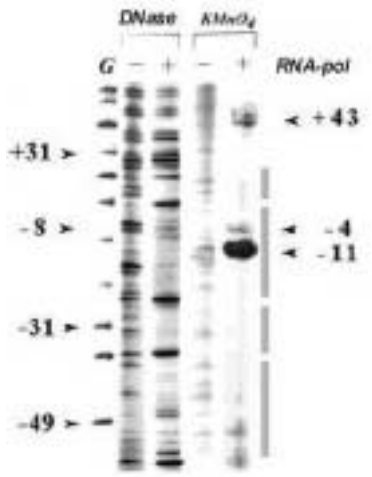


Figure 1. Nucleotide sequence of the promoter T7D promoter including early transcribed region up to position +51. Repeated elements CTTTAG are underlined.

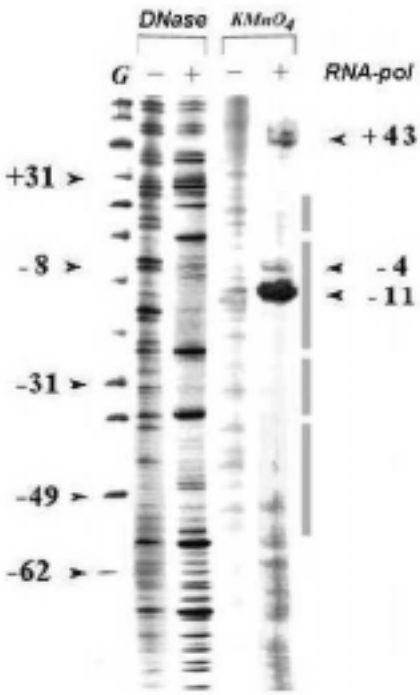


Figure 2. Mapping of structural deformations in the transcription complex formed by *E.coli* RNA polymerase with T7D promoter. Promoter contact area, interacting with the enzyme is marked by vertical bar. The left lane represents G-sequencing ladder.

Acknowledgments

These studies are supported by the Russian Foundation for basic research (grant 00-04-48132) and Havemann scholarship within the frame of "Natural Scientists Initiative "Responsibility for Peace".

References

1. Belyaeva T., Griffiths L., Minchin S., Cole J. and Busby S. (1993) The *Escherichia coli* cysG promoter belongs to the "extended –10" class of bacterial promoters. *Biochem.J.* 296, 851-857.
2. Burns H., Belyaeva T., Busby S. and Minchin S. (1996) Temperature-dependence of open complex formation at two *Escherichia coli* promoters with extended –10 sequences. *Biochem.J.* 317, 305-311.
3. Jones S., van Heyningen P., Berman H. and Thornton J.M. (1999) Protein-DNA Interactions: a structural analysis, *J.Mol.Biol*., 287, 877-896.
4. Khomyakova E.B., Petrova M.V., Minyat E.E. and Ivanov V.I. (1998) Slippedloop structure of DNA: a specific nucleotide sequence forms only one unique conformer. *FEBS Letters* 422, 265-268.
5. Masulis I.S., Chasov V.V., Kostyanicina E.G. and Ozoline, O.N. (1998) Some aspects of protein-DNA recognition in transcription initiation. Molekulyarnaya biologia, 32, 598-602
6. Newlands J., Ros W., Gosnik K.K. and Gourse R.L. (1991) Factor-independent of *Escherichia coli* rRNA transcription. Characterization of complexes of *rrnB* P1 promoters containing or lacking the upstream activator region with *Escherichia coli* RNA polymerase. *J.Mol.Biol.* 220, 569-583.
7. Ozoline O.N. and Tsyganov M.A. (1995) Structure of open promoter complexes of *Escherichia coli* RNA polymerase as revealed by the DNAse1 footprinting technique: compilation analisys. *Nucleic Acids Res*. 23, 4533-4541.
8. Pearson C.E., Eihler E.E., Lerenzetti D., Kramer S.F., Zoghbi H.Y., Nelson D.L. and Sinden R.R. (1998) Interruptions of the triplet repeats of SCA1 and FRAXA reduce the propensity and complexity of slipped strand DNA (S-DNA) formation. *Biochemistry*, 37, 2701-2708.
9. Zaychikov E., Denissova L., Meyer T, Gotte M. and Heumann H. (1997) Influence of Mg^{2+} and temperature on formation of the transcription bubble. *J. Biol. Chem.* 272, 2259-2267.

TRRDEXTR: COMPUTER PROGRAM FOR EXTRACTION OF REGULATORY SEQUENCES DESCRIBED IN TRRD

Kosarev P.S.

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

Keywords: TRRD, regulatory region, transcription factor binding sites, extraction of sequences

Results:

A program is developed for preparing the samples of regulatory elements, for example, transcription factor binding sites (TFBS), promoters, etc. By applying an information on regulatory elements stored in TRRD, the program extracts the sequences corresponding to this information from the EMBL databank.

Introduction

Gene expression regulation is governed at many levels, in particular, regulation of transcription initiation is one of its major stages. Into the process of gene transcription regulation, protein factors are involved (i.e., RNA-polymerase, transcription factors). By interaction of these factors to each other and to DNA regions, they modulate transcription initiation. Information about the participants of this process (transcription factors, their binding sites, composite elements, promoters, enhancers, etc.) and conditions needed for its realization (expression patterns) is accumulated in TRRD database containing hierarchical description of regulatory regions. The scheme of representation the data in TRRD database is shown in Fig. 1. However, TRRD does not contain nucleotide sequences of regulatory elements (except the core sequences for TFBS). With this respect, generating of samples of such sequences is a very important task.

```
[description of a gene]
{features of gene expression in different tissues and cells, that is, the pattern of expression  (supplied by references to regulatory units and sites)
}
{ regulatory region
{  regulatory unit (promoter, enhancer, silencer)
                        {
                        {  composite element  }
                                {  TF binding site
                                {  transcription factor  }
                        }
}
}
}
```

Figure 1. The Structure of blocks used for description of expression patterns and regulatory units in TRRD. The square brackets mean that the block is represented in the entry once, whereas the braces denote that it may be repeated more than once.

For analysis of TFBS properties and recognition of potential TFBS in regulatory regions, the consensuses and weight matrices describing TFBS are often used. However, such form of TFBS representation has some disadvantages, in particular, it does not take into account correlation between nucleotides. Functional sequences of TFBS often possess properties, which are absent in the false TFBS, recognized by weight matrix method within exon sequences (Babenko *et.al.*, 1999). That is why, it is of importance to use the sequences of particular transcription factor binding sites with the flanking regions for analysis, context study, and recognition of potential TFBS within novel sequences.

Results and discussion

We have developed a seria of programs that enable to extract the samples of regulatory sequences (TFBS, promoters, etc.) on the base of their description in TRRD with subsequent usage of the EMBL databank for extraction of corresponding nucleotide sequences. The system developed for extraction of nucleotide sequences of regulatory elements is implemented in the Perl language and it interacts with the TRRD and EMBL databases, which are represented in the form of textual files (Fig. 2).

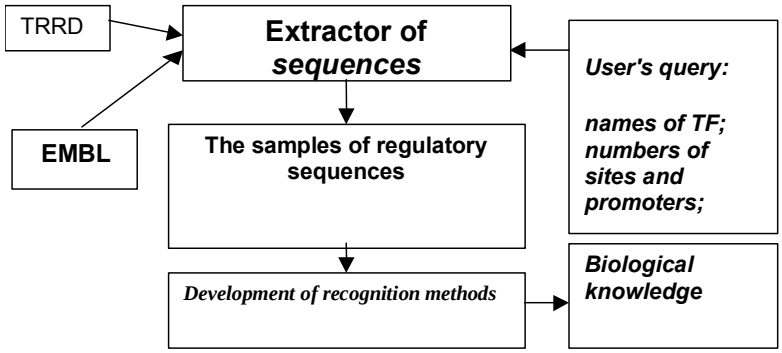


Figure 2. Schematic representation of an extractor of nucleotide sequences of regulatory elements and their automated analysis.

The system is able to process the following types of queries:

Extraction of sites with the flanking regions for transcription factors according to the name of TF given by a user or its synonyms;

Extraction of sites according to their identification numbers in TRRD (these numbers are indicated by experts filling the database);

Extraction of sequences of promoters and enhancers for the pre-ordered groups of genes represented in TRRD: Erythroid-Specific Regulated Genes, Endocrine System Genes, Glucocorticoid-Regulated Genes, Heat Shock-Induced Genes, Lipid Metabolism Genes, and Interferon-Regulated Genes.

The samples of regulatory elements sequences obtained in a result of such extraction, may be represented in several formats convenient for subsequent data processing, including the format adopted in the SAMPLES database implemented for the storage of samples of various functional sequences (Vorobiev *et.al.*, 1998). The examples of the program output are given in Fig. 3.

A)

```
ID   sre_120_006; DNA
AC   sre_120_006
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
OC   Mammalia; Eutheria; Primates; Catarrhini; Hominidae;
OC   Homo.
DR   TRRDSITES; S2091; ; 4.2;
DR   TRRDGENES; ; Hs:SS; 4.2;
DR   EMBL; U18994; HS189941; ; join(1244..1363)
CC   NG squalene synthase
FT   {0,0} [53;68] direct; EXP
FT   SQ = gttATCACGCCAGtct
SQ   caccaatccc gctcgtcggc ctctttctcg gcctccaatg agcttctaga
     gtgttATCAC GCCAGtctcc ttccgcgact gattggccgg ggtcttccta
     gtgtgagcgg ccctggccaa
//
```

B)

```
ID   es_250_12; DNA
AC   es_250_12
DE   cytochrom P450 11 beta hydroxylase gene
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata;
OC   Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
OC   Murinae; Rattus.
DR   EMBL; D14086; RNCYPBA1; ; join(293..542)
CC   ST(EMBL/GENBANK) 493
CC   DR TRRDGENES; A00587; Rn:CYP11B1; 4.2;
FT   {0,0} [1;250]; EXP
SQ   taaagtagag tctgcaccct cccacccacc agcaggcatt gcagaggtag
     gaaaagggag aaagcctcta cctccagaag aaccatcagc tcagtataca
     tttctagggc aagtccaggg acacatcctg cagtgacatt atcagtcagc
     gatttatatc ctcaagacaa gataaaaggc cacggactaa acacaggaag
     agaggaggat ggcaatggct ctcagggtga cagcagatgt gtggctggca
//
```

Figure 3. The program output for transcription factor binding site, by the example of SRE (A) , and for promoter, by the example of a gene referring to endocrine system (B).

By means of the system developed, we have made an extraction of about 60 samples of TFBS and promoters. Some of them are listed in Table 1.

Table 1. List of promoters and transcription factors.

Name of TF or a group of promoters	The length of sequences with the flanking regions, N of nucleotides	Number of sequences
Irf	100	20
Isre	100	35
Stat	100	22
Sre	120	20
Gata1	120	52
c-ebp	>200	52
e2f	>200	21
Myod	>200	10
Sp-1	>200	90
Thr-like family	>200	64
Promoters of endocrine system (ES) genes	250	58

In what follows, the system will be further developed in such a way that the data from TRRD could be represented as the objects corresponding to hierarchical organization of regulatory regions in eukaryotic genes. A possibility will appear that provides making complex queries to many fields of TRRD. This will permit to extract different sorts of samples of regulatory sequences. The system will be available via the Internet.

Acknowledgments

The author is grateful to E.V. Ignatieva, O.A. Podkolodnaya, E.A. Ananko for essential and generous help during the work; to N.A. Kolchanov for scientific supervision and fruitful discussions; to G.V. Orlova for translation of the paper into English.

References

1. Babenko V.N., Kosarev P.S., Vishnevsky O.V., Levitsky V.G., Basin V.V. and Frolov A.S. Investigating extended regulatory regions of genomic DNA sequences. // Bioinformatics, 1999, V.15, Nos 7/8, P. 644-653.
2. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busigina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., Overton K. Transcription Regulatory Regions Database (TRRD):its status in 2000. // Nucleic Acids Research, 2000, 28(1), 298-301.
3. Vorobiev D.G., Ponomarenko J.V., Podkolodnaya O.A Samples and Aligned: databases for functional site sequences. // Proc. I Intern. Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, Russia, 1998, p. 58-61.

PERIODIC PATTERNS IN SEQUENCE ORGANIZATION OF REPLICATION ORIGIN OF *ESCHERICHIA COLI* K-12 CHROMOSOME

Kravatskaya G.I., Esipova N.G.

V.A. Engelhardt Institute of Molecular Biology, Moscow, Russia
e-mail: GK@imb.imb.ac.ru, nge@imb.imb.ac.ru
*Corresponding author

Keywords: origin of chromosome replication, DNA unwinding, periodicity, matrix Fourier analysis

Resume

Motivation:

The process of initiation of DNA replication is one of the most important and insufficiently known in the cell. This process starts at special DNA sequences called replication origins. E.coli chromosomal replication during normal growth initiates bidirectionally at unique sequence (*oriC*). In order to reveal periodic regularities in the primary structure of *oriC* and analyze their role in the process of replication initiation, we applied matrix Fourier analysis.

Results:

oriC Fourier spectra were obtained and compared with that ones of other regions of *E.coli* complete genome. Using matrix Fourier analysis with sliding window technique several sites resembling *oriC* were revealed. Most of them are coincide with the sites of replication initiation in *E.coli* stable DNA replication mutants. The method applied can be useful for analysis of other complete genomes and prediction of the sites of possible replication initiation.

Introduction

The replication of the *E.coli* chromosome is normally initiated at the *oriC* site, the origin of replication. A sequence of 245 base-pairs (*oriC*) in the replication origin of the *E.coli* K-12 chromosome has been shown to provide all the information essential for initiation of bidirectional replication [Asada et al., 1982]. It is known that *oriC* sequence is extremely saturated with direct and inverted repeats [Meijer et al., 1979]. In this work we attempted to reveal another (periodic) regularities in the *oriC* nucleotide sequence. Discovering of periodicities in the DNA primary structure is important for understanding of regularities of higher order structures formation and stability.

Methods and algorithms

The Fourier transformation of nucleotide sequence is performed as described in [Makeev and Tumanyan, 1996]. The program applied was PERF [Makeev et al., 1996].

Implementation and results

Periodicities in the dispositions of nucleotides and dinucleotides in the origin of chromosome replication *oriC* from *E.coli* were studied by means of matrix Fourier analysis. Peaks corresponding to the periods T=2, 17, 93-98 nucleotides are the most high in the Fourier spectrum of *oriC* (Fig.1). Peaks corresponding to the periods T= 3, 11, 19, 13, 24, 27, 28, 41, 79-81 nucleotides are also prominent, but not so high. The difference between *oriC* Fourier spectrum and that ones of adjacent to *oriC* regions are demonstrated (Fig.2, 3).

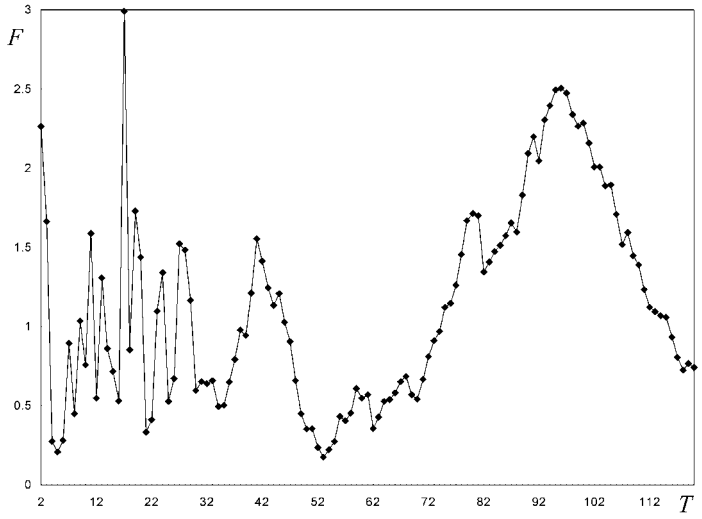


Figure 1. Fourier spectrum of *oriC* in terms of A,C,G,T nucleotide occurrences. T- the length of the period, F- corresponding to the period of T spectral power.

We have also demonstrated that *oriC* contains several regions with different periodic organization of nucleotide occurences. We connected this result with the disposition on *oriC* of binding sites of initiator protein DnaA and regulatory proteins FIS and IHF [Woelker and Messer, 1993].

Matrix Fourier analysis of *E.coli* genome with sliding window (step = 100 nucleotides, the length of the window = 245 nucleotides) technique reveals that only 10 regions of *E.coli* genome have Fourier spectra resembling the Fourier spectrum of *oriC* in the sense of prominent peaks corresponding to the periods of T=2, 17, 93-98.

Discussion

The periodicity, corresponding to B-DNA pitch (10.5), is negligible in the Fourier spectrum of *oriC* in contrast to the Fourier spectra of flanking regions (Fig.2). The main periodicities of the oriC spectrum are not multiples of the B-DNA sugar-phosphate backbone period, that destabilizes DNA at *oriC* and contributes to the formation of a structure called a replication bubble. We suppose that the presence of strong periodicities destabilizing the B-form of DNA also contributes to the spontaneous unwinding [Polaczek et al., 1998] of DNA in *oriC*.

In stable DNA replication sdr/rnh mutants of *E.coli*, initiation of replication occurs in the absence of the normal origin of replication, *oriC* [de Massy et al., 1984]. There are at least four fixed sites or regions of the sdrA Δ*oriC* chromosome from which DNA replication can be initiated in the absence of the oriC sequence. Most of the regions revealed by our method are the sites of replication initiation in E.coli stable DNA replication (sdrA/rnh) mutants. Two of the sites revealed are novel sites. Probably, these sites are also functional but normally repressed. Our results suggest that the method applied can be useful for analysis of other complete genomes and prediction of the sites of possible replication initiation.

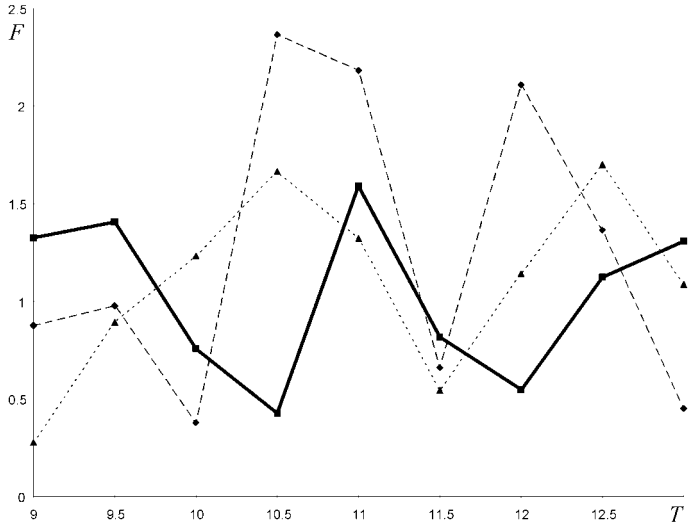


Figure 2. Fragments of the Fourier spectra of oriC (solid bold line) and that ones of the regions (dotted line), which flank oriC (in terms of A,C,G,T nucleotides occurences). T- the length of the period, F- corresponding to the period of T spectral power.

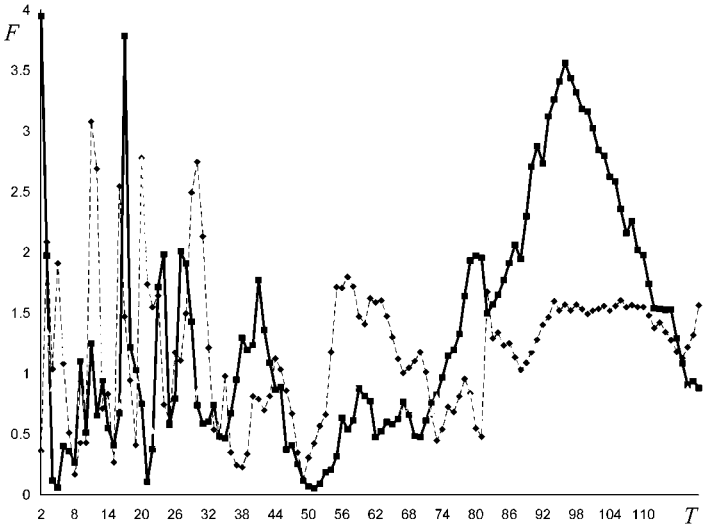


Figure 3. Fourier spectra of *oriC* (solid bold line) and flanking regions (dotted line) in terms of A and T nucleotides occurrences. T- the length of the period, F- corresponding to the period of T spectral power

Acknowlegements

This work was supported by the grant N 00-04-48351 from Russian Foundation of Basic Research (RFBR).

References

1. Asada K., Sugimoto K., Oka A., Takanami M., Hirota Y., (1982) Structure of replication origin of the *Escherichia coli* K-12 chromosome: the presence of spacer sequences in the ori region carrying information for autonomous replication *Nucleic Acids Res* 10(12), 3745-54.
2. Makeev V.Ju., Tumanyan V.G. (1996) Search of periodicities in primary structure of biopolymers: a general Fourier approach. CABIOS 12, 49-54.
3. Makeev V.Ju., Frank G.K. , Tumanyan V.G. (1996) Statistics of periodic patterns in the sequences of human introns. Biophysics, 41, 1, 263-268.
4. de Massy B., Fayet O., Kogoma T. (1984). Multiple origin usage for DNA replication in sdrA(rnh) mutants of Escherichia coli K-12. Initiation in the absence of oriC. JMolBiol 178(2), 227-36.
5. Meijer M., Beck E., Hansen F.G., Bergmans H.E.N., Messer W., Meyenburg K., Schaller H., (1979) Nucleotide sequence of the origin of replication of the *Escherichia coli* K-12 chromosome Proc. Natl. Acad. Sci. USA. 76, 2, 580-584.
6. Polaczek P., Kwan K., Campbell J.L. (1998) Unwinding of the *Escherichia coli* origin of replication (oriC) can occur in the absence of initiation proteins but is stabilized by DnaA and histone-like proteins IHF or HU. *Plasmid* 39 (1), 77-83.
7. Woelker B., Messer W., (1993) The structure of the initiation complex at the replication origin, oriC, of *Escherichia coli*. Nucleic Acids Res. 21, 22, 5025-5033.

CHARACTERISTIC MODULAR PROMOTER STRUCTURE AND ITS APPLICATION TO DEVELOPMENT OF RECOGNITION PROGRAM SOFTWARE

Levitsky V.G. and Katokhin A.V.

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

Keywords: recognition, promoter, discriminant analysis, modular structure

Resume

Motivation:

Recognition of promoter gene regions is very important procedure for detection of genes and their structure. However, current methods for promoter recognition are not always supplied by high accuracy of recognition.

Results:

The novel algorithm for promoter recognition is developed. It is based on dissection of promoter region into non-overlapping parts. In what follows, recognition function is constructed by discriminant analysis according to dinucleotide frequencies in each region. The method developed is implemented for recognition of promoters in *Drosophila melanogaster*.

Introduction

Advances in genome sequencing in eukaryotes cause the necessity to develop computer methods for gene recognition. Promoter is a key element of DNA structure that is obligatory for transcription. It consists of a seria of regulatory elements – short DNA motifs, which serve as the sites binding proteins (transcription factors, TFs) [Pedersen et al., 1999]. Among the most significant are such elements as TATA-box, initiator (Inr-element) [Bucher, 1990], and DPE-element [Burke and Kadonaga, 1996]. For promoter recognition, the following methods are used: Markov chain models [Audic and Claverie, 1997], neuron networks [Knudsen, 1999], and discriminant analysis [Zhang, 1998]. In the present paper, we give a description of the novel method for prediction of promoters in *Drosophila melanogaster*.

Initial data and methods

Promoter sequences of *Drosophila melanogaster*, within the range [-300; +100] referring to start site of transcription (SST), were extracted from the DPD database [Arkhipova, 1995]. For classification of promoters, weight matrix for the TATA-box was applied [Bucher, 1990]. All promoters were divided into three subsets: TATA-containing (TATA+), TATA-intermediate, and TATA-less (TATA-) (Table 1). The goal of this dissection is to increase an accuracy of promoter recognition in each distinct subset. An intermediate group was dissected for more strict manipulations with the sets of TATA+ and TATA- promoters.

Table 1. Classification of promoters according to weight matrix value for the TATA-box.

Name of a set	Weight matrix value, Score	Number of sequences
TATA+	Score ≥ -5	60
Intermediate	-9 ≥ Score >-5	35
TATA-	Score < -9	108

Discriminant recognition function operates with two sets of sequences: the positive one contains promoters, whereas the negative one – random sequences, generated with conservation of the same nucleotide content as in the set of promoters. As independent variables of discriminant analysis, we used the nucleotide frequencies for particular regions of promoter sequences. On the basis of Monte-Carlo method, we have developed an iterative program, which determines an optimal dissection of the whole promoter region (400 bp) into 12 non-overlapping parts (Figure 1). The goal of this program is to determine the boundaries of each region. We were searching for such dissection that maximizes the value R^2 (Mahalanobis distance) calculated by discriminant analysis. The growth of this value may be interpreted as the more distinct separation of positive and negative sets of sequences. To reveal the final dissection of the whole promoter region into partitions, we have used an iterative procedure, which is free of pre-set limitations on position of boundaries between the separate segments. Besides, by the following calculation of the exact discriminant function, we take into account mutual correlation frequencies between the segments. We applied the following formula of recognition function:

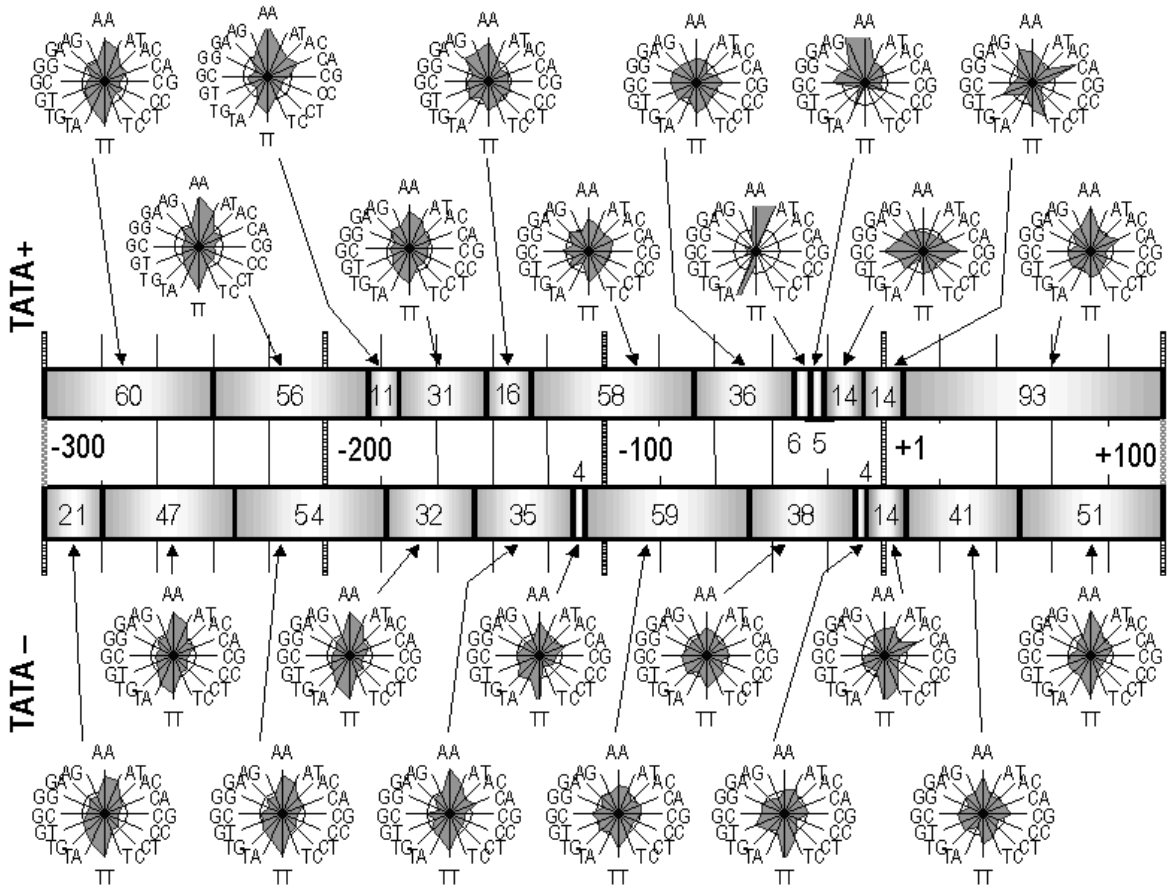


Figure 1. Dissection of promoter region [-300;+100] into the sections calculated for promoters of TATA+ and TATA- types. Locations and distributions of dinucleotide frequencies for each section are shown too.

$$\varphi(f) = \frac{1}{R^2} * \sum_{j=1}^{N} \{[f_j - (1/2) * (\overline{f_j^{(2)}} + \overline{f_j^{(1)}})] * d_j\}.$$

Here f_i denotes the frequency of a dinucleotide, N – total number of variables. The values d_j and R^2 are calculated by the formulas:

$$d_j = \sum_{k=1}^{N} S_{j,k}^{-1} * [\overline{f_k^{(2)}} - \overline{f_k^{(1)}}],$$

$$R^2 = \sum_{k=1}^{N} \sum_{j=1}^{N} \{[\overline{f_j^{(2)}} - \overline{f_j^{(1)}}] * S_{j,k}^{-1} * [\overline{f_k^{(2)}} - \overline{f_k^{(1)}}]\}.$$

Here by S^{-1} is denoted a reverse matrix for united covariation matrix; $S = S^{(1)} + S^{(2)}$; $S^{(1)}$ and $S^{(2)}$ are covariation matrices for the positive and negative sets of sequences, respectively. If the value of recognition function is close to 1, then recognition is significant. The value equaling to –1 evidences about the similarity of the sequence considered to the random sequences.

In order to evaluate an accuracy of promoter recognition, we used the standard statistical measures. Let us denote TP and TN as the numbers of properly and improperly predicted promoters, and FP and FN - as the numbers of correctly and incorrectly predicted non-promoters. By applying these denotations, we have determined false negative estimates E_1 (under-recognition, that is, some promoters are not recognized) and false positive estimates E_2 (over-recognition, or some non-promoters were recognized as promoters). In addition, we used the following measures for estimating recognition accuracy:

$$Sn = 1 - E_1 = \frac{TP}{TP + FN}, \text{ sensitivity,}$$

$$Sp = 1 - E_2 = \frac{TN}{TN + FP}$$, specificity.

The general accuracy evaluation of the method suggested is based on calculation of correlation coefficient CC:

$$CC = \frac{TP * TN - FN * FP}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}}.$$

Results and discussion

Optimal dissections of the whole promoter region and corresponding frequency distributions are illustrated in Figure 1. Polygon diagram (Figure 1) is convenient and illustrative tool for representation of a complex structure of different types of promoters. TATA+ and TATA – promoters are described by different inherent structure of regulatory modules. The TATA+ promoter set is characterized by the pronounced signal in the TATA-box region. The canonical consensus TATAAA was found to be dissected into two sections (with the lengths equaling to 6 and 5 bp, respectively). For TATA- promoters, there is an extensive section, with the length of 38 bp, which is not marked by contrast peculiarities of dinucleotide content. In addition, for the TATA+ promoter set, a signal was detected within the region of the Inr-element. Within the limits of the section with the length of 14 bp, which intersects with transcription start, the most frequent are CA, AG, GT, and TC dinucleotides. In the region downstream transcription start, the distinct context signals are absent. This could mean that in case effective TATA-box is present, then additional regulatory elements downstream TSS are not necessary. For the TATA- promoter set, the context signals corresponding to Inr- and DPE elements are typical. In the region of the Inr-element, the section with the length of 14 bp is present, the most typical being CA, TT, and TC dinucleotides. In this region, promoters of TATA+ and TATA- types have the similar structure. The most notable feature of the TATA- promoters is the region, with the length of 41 bp, located upstream TSS. Notably, among all the subsets of TATA+ and TATA- promoters, this region is characterized by the maximal frequency of CG dinucleotides. Exactly this dinucleotide is known to be most characteristic for the DPE element [Burke and Kadonaga, 1996; 1997].

Accuracy of promoter recognition was evaluated by cross-testing of promoter sets against the random sequence set and genome sequence set containing exons and introns (Tables 2 and 3). The sets of exons and introns were extracted from the EMBL databank. Evaluation of accuracy obtained for the set of intermediate promoters was the worst. We suppose this fact is explained by non-homogeneity of this promoter group.

Table 2. Recognition accuracy estimated by means of random sequences.

Name of a set	Sensitivity (Sn)	Specificity (Sp)	Correlation coefficient (CC)
TATA+	0.79	0.97	0.74
TATA-intermediate	0.72	0.96	0.62
TATA-	0.81	0.92	0.71

Table 3. Evaluation of accuracy of prediction calculated for genome sequences of exons and introns.

Name of a set	Sensitivity (Sn)	Specificity (Sp)	Correlation coefficient (CC)
TATA+, introns	0.63	0.99	0.70
TATA+, exons	0.63	0.995	0.72
TATA-, introns	0.43	0.95	0.40
TATA-, exons	0.52	0.98	0.53

Comparison of our recognition group with the other programs demonstrates that the quality of our recognition is at least not worse than the others. By some parameters, our method gives even better results. However, it should be noted that such comparisons are not very reliable, because initial data and the data used for preparation of the training and control sets may vary in different programs. Namely, only in a single publication [Ohler and Reese, 1998], promoters of exactly *Drosophila melanogaster* species were used as the training set for recognition program.

Detected modular structure of TATA+ and TATA- promoters is likely related to characteristic inherent structure of these types of promoters. The usage of the most useful dissection of the samples of promoters and non-promoters in the process of recognition function construction has enabled us to take into account fine features of promoter region structure. Besides, by considering interrelated correlations of dinucleotide frequencies in the sections, it is possible to consider weakly expressed or inaccurately localized context signals (e.g., DPE-element in TATA- promoters).

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants Nos 98-07-90126, 98-07-91078, 99-07-90203) and by the grant for Young Scientists by Siberian Branch of RAS. The authors are grateful to Galina Orlova for translation of the manuscript into English.

References

1. Arkhipova I.R., Promoter elements in Drosophila melanogaster revealed by sequence analysis. *Genetics,* 1995, 139, 1359-1369.
2. Audic S., Claverie J.M., Detection of eukaryotic promoters using Markov transition matrices. *Comput. Chem.*, 1997, 21, 223-227.
3. Bucher P., Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.*, 1990, 212, 563-578.
4. Burke T.W., Kadonaga J.T., Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.,* 1996, 10, 11-24.
5. Burke T.W., Kadonaga J.T., The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. *Genes Dev.,* 1997, 11, 3020-3031.
6. Levitsky V.G., Katokhin A.V., Kolchanov N.A. Inherent modular promoter structure and its application for recognition tools development. *Computational technologies (Novosibirsk)*, 2000, 5, special issue, 41-47.
7. Knudsen S., Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics*, 1999, 15, 356-361.
8. Ohler, U. and Reese, M., Detection of eukaryotic promoter regions using polygrams. In R. Hofestädt, "Molekulare Bioinformatik", Shaker, Aachen, 1998, 89-100.
9. Pedersen A.G., Baldi P., Chauvin Y., Brunak S., The biology of eukaryotic promoter prediction--a review. *Comput. Chem.*, 1999, 23, 191-207.
10. Zhang M.Q., A discrimination study of human core-promoters. *Pac Symp Biocomput* 1998, 240-251.

NUCLEOSOME ORGANIZATION OF CHROMATIN IN EUKARYOTIC GENES AND STRUCTURE-FUNCTIONAL GENOME REGIONS

Levitsky V.G., Kolchanov N.A.

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

Keywords: nucleosome positioning, chromatin structure, nucleosomal site recognition, discriminant analysis

Resume

Motivation:

In order to analyze a huge massive of poorly studied genome sequences, it is necessary to develop novel computer tools. In particular, since nucleosome packaging is very important for proper DNA functioning, detection of specific nucleosome packaging features within genes or particular genome regions could be useful for studying the structure and function of these regions.

Results:

A program was developed for recognition of nucleosome site positioning along DNA sequences in eukaryotes. An analysis of recognition function distribution was performed for different functional types of genome DNA sequences. Several common features of structural organization of genome and some genes were revealed, being related to nucleosome positioning.

Availability:

The program for nucleosome sites recognition is included into the GeneExpress system, section DNA nucleosomal organization, http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/.

Introduction

Chromatin organization into nucleosomes is the basal process of DNA packaging in eukaryotes. Moreover, nucleosome positioning is one of the factors controlling gene transcription. The study of nucleosome positioning within DNA has revealed that in many cases, the preferential nucleosome positioning is observed in particular DNA sites. For instance, a regularity in some di- and trinucleotides distribution in nucleosomal DNA was observed [Satchwell et al., 1986; Ioshikhes et al., 1999; Stein and Bina, 1999]. Besides, the preferential nucleosome positioning in particular sites is related to some conformational DNA features [Levitsky et al., 1999]. Such features may be produced in different ways, that is, differentially selected DNA context is responsible for nucleosome positioning [Widlund et al., 1997]. Generalization of many experimental and theoretical data on nucleosome positioning in genome DNA enabled to conclude that there exists a special weak context code of nucleosome positioning [Trifonov, 1997]. The peculiarities of this code are as follows: (1) the code is of extremely degenerate nature (very differing DNA sequences are capable to interact with histone octamer with subsequent nucleosome organization); (2) contextual signals are very weak; (3) context signals within the region of DNA interaction with histone octamer have no clear localization.

Thus, histone octamer is able to interact with very differing DNA-sites. The obligatory signals are absent in this code, so the positioning of the core octamer in the concrete DNA-site is based on specific for this site combination of signals localized in specifically determined positions. Notably, formation of concrete nucleosome type is supported by the definite subset of signals out of their potentially large variability. By using these very peculiarities, we have developed a method for recognition of nucleosome site positioning on the base of discriminant analysis [Levitsky et al., 2000]. In this case, the dinucleotide concentrations in different regions of nucleosome site are considered as the signals. By means of this method, we have analyzed the samples of genome sequences in eukaryotes. It was found that nucleosome positioning along the genes is characterized by a variety of common features. Out interest to the studying of characteristic features of nucleosomal DNA is attracted also by the fact that currently, more and more peculiarities of nucleosome positioning are being found in the neighborhoods of transcription start site [Ioshikhes et al., 1999; Pedersen et al., 1999].

Initial data and methods

To construct nucleosome sites recognition method, we have used two sets of nucleotide sequences: the set of nucleosomal sites [Ioshikhes & Trifonov, 1993] and the random sequences set. Then by discriminant analysis,

dinucleotide frequencies in particular regions of a nucleosome site are taken into account [Levitsky et al., 2000]. Calculation of the function is made by the formula:

$$\varphi(f) = \frac{\sum_{k=1}^{N}\sum_{j=1}^{N}\{[f_j - (1/2)*(\overline{f_j^{(2)}} + \overline{f_j^{(1)}})]*S_{j,k}^{-1}*[\overline{f_k^{(2)}} - \overline{f_k^{(1)}}]\}}{\sum_{k=1}^{N}\sum_{j=1}^{N}\{[\overline{f_j^{(2)}} - \overline{f_j^{(1)}}]*S_{j,k}^{-1}*[\overline{f_k^{(2)}} - \overline{f_k^{(1)}}]\}}.$$

Here we make the summation with respect to the set of dinucleotide frequencies $\{f_k\}$; indices (1) and (2) denote the number of a set, S^{-1} is a reverse matrix for the united covariation matrix S, $S = S^{(1)} + S^{(2)}$; $S^{(1)}$ and $S^{(2)}$ are covariation matrices for two sets of sequences. The proximity of recognition function to 1 means the better ability of this sequence to nucleosome positioning.

In order to study nucleosome DNA packaging near transcription start sites, we have isolated promoter sequences from the EPD database [Perier et al., 2000]. For analysis of peculiarities of nucleosome positioning at the borders between exons and introns, we have compiled the samples of human splicing sites from the EMBL data bank.

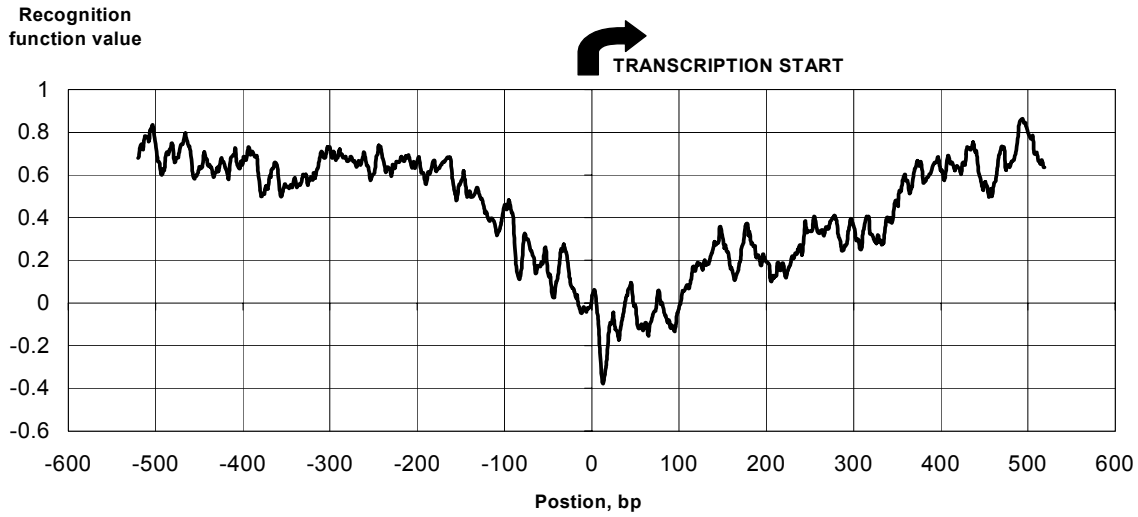


Figure 1. Nucleosome site recognition profile for human genes promoter regions.

Results and discussion

By means of the nucleosome site recognition program, we have analyzed the samples of promoters, splicing site regions, introns and exons. The nucleosome recognition function profile calculated for the human promoters is shown in Figure 1.

As can be seen from the Figure, the region nearby transcription start is characterized by the minimal recognition function values. The clearly distinguished minimum in the transcription start region may indicate that nucleosome DNA packaging in this region is either weak or even absent. This situation favors to the easy access of the basal transcription complex proteins to DNA, which is necessary for quick transcription initiation. The result obtained provides evidence about priority of nucleosome code for processing of eukaryotic genes transcription. This result, together with the recent studies revealing common regularities of nucleosome positioning in the neighborhoods of transcription start sites [Baldi et al., 1998, Ioshikhes et al., 1999], causes the necessity of more detailed analysis of nucleosomal DNA organization within promoter gene region. One of our recent publications is devoted to comprehensive consideration of the problem on relationships between nucleosome positioning nearby transcription start sites and the pattern of gene expression [Levitsky & Podkolodnaya, 2000].

The question on evolutionary motivation of the origin of introns is one of the intriguing problems in molecular biology. The recognition function profile calculated for donor and acceptor splicing sites is shown in Figure 2. The transition from exon to intron is marked by increase in recognition function value, and its local maxims are located in introns nearby the boundary with exons. Thus, exons have the low potential for nucleosome positioning in comparison with introns. The possible reasoning of this fact may be the functional loading of

exons by genetical code. That is why localization of effective nucleosome positioning sites is difficult in exons. As for introns, they are not loaded by genetical code and may easily perceive the signals of nucleosome positioning. An increased ability for nucleosome positioning that was detected in introns near by the boundaries with exons may be necessary for providing packaging of DNA of the whole transcription unit in regulatory chromatin structures. Detection of nucleosome positioning sites in introns, in the neighborhoods of splicing sites, supports a hypothesis suggested earlier about evolutionary motivation of the origin of introns. It was supposed that if the gene lacks the signal for nucleosome positioning due to restrictions set by the primary protein structure, then the problem of its DNA packaging can be solved by insertion of intron containing effective nucleosome positioning signal into the corresponding position [Solovyev & Kolchanov, 1985; Csordas, 1989; Denisov et al., 1997].

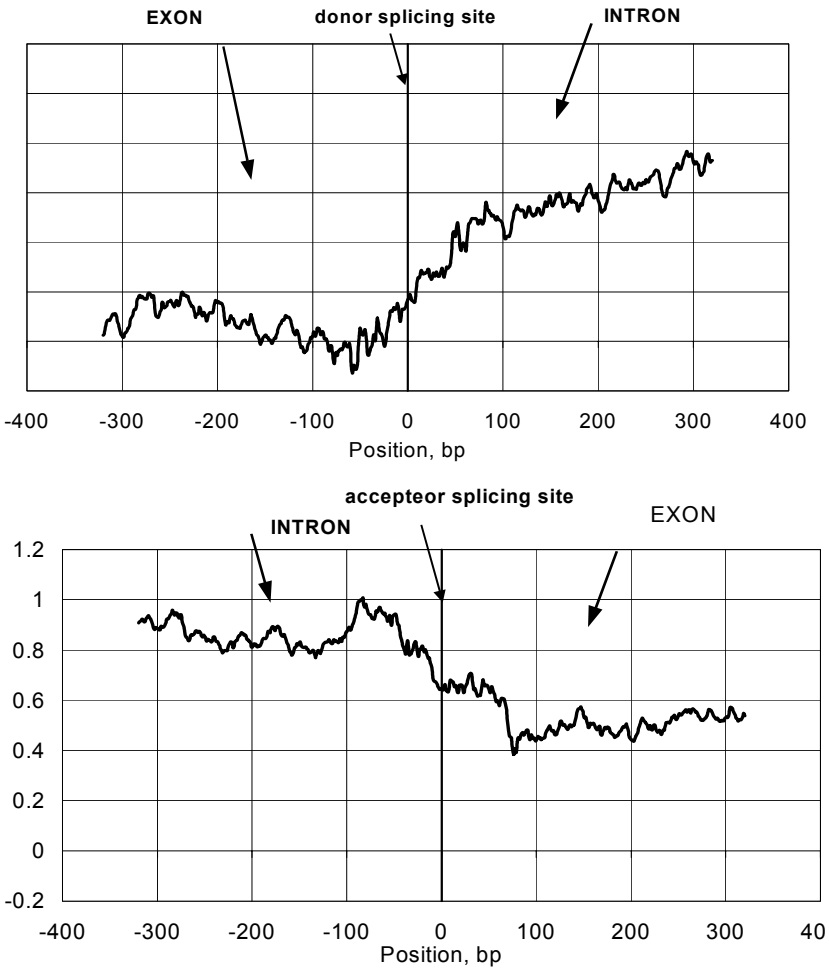


Figure 2. Profiles of the function for nucleosome recognition within the human splicing sites.

In Figure 3, the recognition function profile of nucleosome site positioning in the regions of mobile element P1 insertions is given. The site sequences of P1 element insertions were extracted from the database of Drosophila Genome Project (Berkeley) (ttp://www.fruitfly.org/sequence/download.html) Notably, the region of insertion in comparison with the flanking regions is characterized by decreased values of the recognition function. This means that P1 element insertion occurs preferentially in the chromatin regions with decreased ability to nucleosome positioning. Really, mobile element insertion into the less densely packaged chromatin regions is more probable than into the regions with increased ability to nucleosome positioning.

Figure 3. Nucleosome site recognition profile for P1 mobile element insertion in *Drosophila melanogaster*.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grants Nos. 98-07-90126, 98-07-91078, 99-07-90203) and Young Scientists Competition of SB RAS. The authors are grateful to G.V. Orlova for translation of the paper into English.

References

1. Baldi P., Chauvin Y., Brunak S., Gorodkin J., Pedersen A.G. Computational applications of DNA structural scales. *Ismb.* 1998, 6, 35-42.
2. Csordas A. A proposal for a possible role of nucleosome positioning in the evolutionary adjustment of introns. *Int J Biochem.*, 1989, 21, 455-461.
3. Denisov D.A., Shpigelman E.S., Trifonov E.N. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. *Gene* 1997, 205, 145-149.
4. Ioshikhes I., Trifonov E.N. Nucleosomal DNA sequence database. *Nucl. Acids. Res.*, 1993, 21, 4857-4859.
5. Ioshikhes I., Trifonov E.N., Zhang M.Q. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc Natl Acad Sci USA,* 1999, 96, 2891-2895.
6. Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Kolchanov N.A., Nucleosomal DNA property database. *Bioinformatics*, 1999, 15, 582-592.
7. Levitsky V.G., Podkolodnaya O.A. *Bioinformatics of Genome Regulation and Structure* (this issue), 2000.
8. Levitsky V.G., Katokhin A.V., Kolchanov N.A. Inherent modular promoter structure and its application for recognition tools development. *Computational technologies (Novosibirsk)*, 2000, 5, special issue, 41-47.
9. Pedersen A.G., Baldi P., Chauvin Y., Brunak S., The biology of eukaryotic promoter prediction--a review. *Comput. Chem.*, 1999, 23, 191-207.
10. Perier R.C., Praz V., Junier T., Bonnard C., Bucher P. The eukaryotic promoter database. *Nucleic Acids Res,* 2000, 28, 302-303.
11. Solovyev V.V., Kolchanov N.A. The eucaryotic genes exon-intron structure can be determined by the nucleosomes organisation of the chromatin and related characteristics of gene expression regulation. *Dokl. Akad. Nauk SSSR,* 1985, 284, 232-237.
12. Satchwell S.C., Drew H.R. and Travers A.A. Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.,* 1986, 191, 659-675.
13. Stein A., Bina M. A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res*., 1999 27, 848-853.
14. Trifonov,E.,N. Genetic level of DNA sequences is determined by superposition of many codes. *Mol Biol (Mosk)* 1997, 31, 759-767.
15. Widlund H.R., Cao H., Simonsson S., Magnusson E., Simonsson T., Nielsen P.E., Kahn J.D., Crothers D.M., Kubista M. Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.,* 1997, 267, 807-817.

ANALYSIS OF RELATIONSHIPS BETWEEN NUCLEOSOME POSITIONING IN PROMOTER REGIONS AND GENE EXPRESSION PATTERN

Levitsky V.G., Podkolodnaya O.A.

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

Keywords: nucleosome positioning, gene expression pattern, “housekeeping” genes, tissue-specific genes

Resume

Motivation:

Rapid growth in numbering of known genes provokes to develop automated tools for their classification and analysis. In the last years, it becomes clear that nucleosome packaging of promoter gene regions is very important for gene functioning. Application of automated computer tools for nucleosome packaging density recognition in nucleosome regions in eukaryotes may be used for studying of the pattern of gene expression.

Results:

An analysis was performed of nucleosome packaging density for human genes differing by expression patterns. The potential of nucleosome packaging DNA for promoters of genes with tissue-specific activity appeared to be essentially higher than that of genes expressed in many tissues or that of “housekeeping” gene promoters. This phenomenon counts in favor that one of the factors regulating gene expression pattern is ability of a gene promoter to nucleosome positioning.

Availability:

The program for nucleosomal sites recognition is included into the GeneExpress system, section DNA nucleosomal organization, http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/.

Introduction

The basal level of DNA packaging is its nucleosomal organization. DNA features determining nucleosome positioning are important both for proper DNA packaging and functioning [Paranjape et al., 1994]. In many cases, nucleosome positioning in promoter region controls the gene activity: positioned nucleosome may inhibit or enhance gene expression [Wolffe, 1994]. Some general features of promoter DNA related to nucleosome positioning [Ioshikhes et al., 1999; Levitsky et al., 1999] were found. Peculiarities of promoter DNA related to nucleosome positioning could be used for development of promoter recognition methods [Pedersen et al., 1999; Levitsky et al., 2000]. With this respect, it seems very interesting to analyze and search for DNA features responsible for nucleosomal sites positioning within genome sequences. Currently, there exist some approaches for revealing context DNA features significant for regular nucleosome formation. Among the examples are detection of periodicity in di- or tri-nucleotides content [Satchwell et al., 1986; Ioshikhes et al., 1996; Stein and Bina, 1999], studying of DNA conformational features [Fitzgerald et al., 1994; Sivolob and Kharpunov, 1995]. For nucleosome sites recognition, we have used the method based on discriminant analysis [Levitsky et al., 2000].

Methods and algorithms

Recognition function is constructed on the basis of analysis of nucleosomal site samples [Ioshikhes and Trifonov, 1993] and random sequences. By the method of discriminant analysis, dinucleotide frequencies were calculated for some nucleosomal sites regions detected in a special way. Calculation of recognition function is produced by the formula:

$$\varphi(f) = \frac{\sum_{k=1}^{N}\sum_{j=1}^{N}\{[f_j - (1/2)*(\overline{f_j^{(2)}} + \overline{f_j^{(1)}})]*S_{j,k}^{-1}*[\overline{f_k^{(2)}} - \overline{f_k^{(1)}}]\}}{\sum_{k=1}^{N}\sum_{j=1}^{N}\{[\overline{f_j^{(2)}} - \overline{f_j^{(1)}}]*S_{j,k}^{-1}*[\overline{f_k^{(2)}} - \overline{f_k^{(1)}}]\}}.$$

Here the summing is made by the set of dinucleotide frequencies $\{f_k\}$, calculated for the definite site regions; by indices (1) and (2), the number of a set is denoted; S^{-1} denotes the reverse matrix for united covariation matrix S; $S = S^{(1)} + S^{(2)}$; $S^{(1)}$ and $S^{(2)}$ are covariation matrices for two sets of sequences. Close correspondence of recognition function to 1 means better ability of the sequence analyzed to nucleosome positioning.

Implementation and results

To analyze the dependency of expression pattern upon DNA packaging density in the regions of 219 human promoters within the interval [-300;+100] relatively transcription start were extracted from the EPD database [Perier et al., 2000]. Classification of promoters according gene expression pattern enabled to detect distinct promoter classes. For this purpose, we have used the databases EPD [Perier et al., 2000], TRRD [Kolchanov et al., 2000], and literature sources. The names and number of promoters are given in Table 1.

Table 1. Classification of human promoters according to expression pattern and distribution of nucleosomal site recognition function values by the classes of promoters differing by expression efficiency. Distribution values are calculated for the region [-50; +1] of promoters.

Name of promoter class	Number of promoters	Average value	Number of function values in the interval (%)	
			[0; 2]	out of [0 ,2]
"Housekeeping' genes	32	-1.48±0.04	11.11%	88.89%
Genes expressed in a wide range of tissues	30	-0.66±0.04	21.77%	78.13%
Tissue-specific genes	141	+0.70±0.007	80.15%	19.85%

We have previously shown that promoter region located exactly prior transcription start site is characterized by decreased values of nucleosomal sites recognition function [Levitsky et al, 1999]. In Figure 1, the distribution is shown of nucleosomal sites recognition function values within the region [-50;+1] relatively transcription start. It could be noted that promoters of "housekeeping" genes and that of widely expressed genes are characterized by the lowest function values.

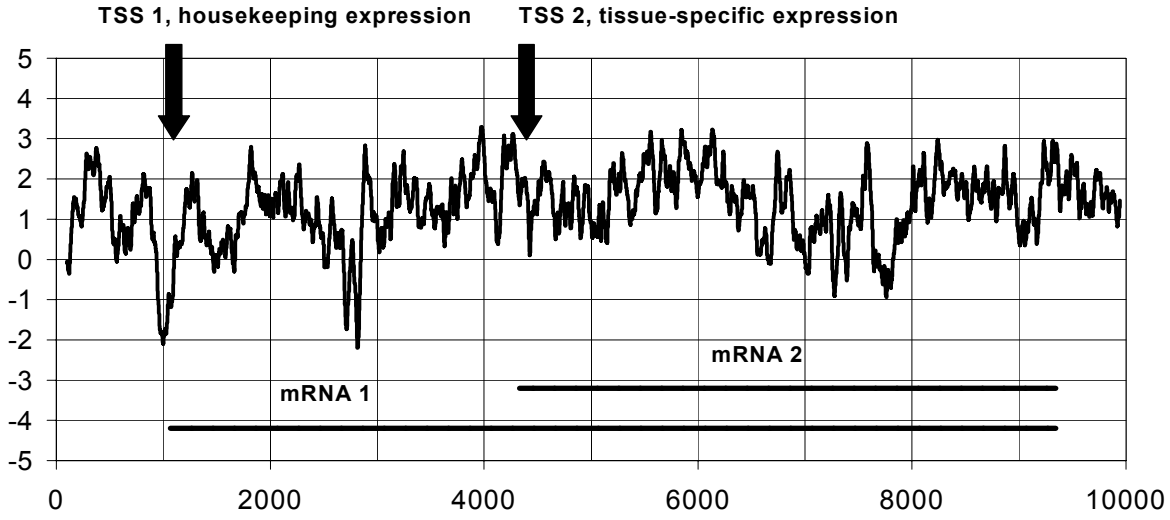


Figure 1. Recognition function profile for nucleosomal sites of the human hydroximethylbilansynthetase gene (AC M95623).

The class of tissue-specific genes promoters are characterized by the visible shift to the right in comparison to promoters of "housekeeping" and widely expressed genes. The brief report of nucleosomal sites recognition function values for various promoter classes is given in Table 1. In the Table, for each class are given the following data: number, mean value, its standard deviation and the share of promoters in the class, such that the recognition function values fall within and outside the interval [0; +2]. The mean function values in the class of tissue-specific gene promoters differ significantly from the mean values of "housekeeping" and widely expressed genes promoters classes according to the Student's criterion with the significance level α>0.99.

Discussion

Comparison of typical nucleosomal sites recognition function profiles for the genes with different expression pattern is demonstrated in Figures 2 and 3. In Figure 2, the gene is shown, such that its expression is controlled by two promoters: the first promoter controls the synthesis of 'housekeeping" product, whereas the second one is tissue-specific. In Figure 3, the profiles for two genes are given: the ubiquitin gene is widely expressed in many tissues, on the contrary, prealbumin gene is tissue-specific. It can be seen that for the genes with tissue-specific pattern of expression, the recognition function values in promoter region are essentially more close to a unit.

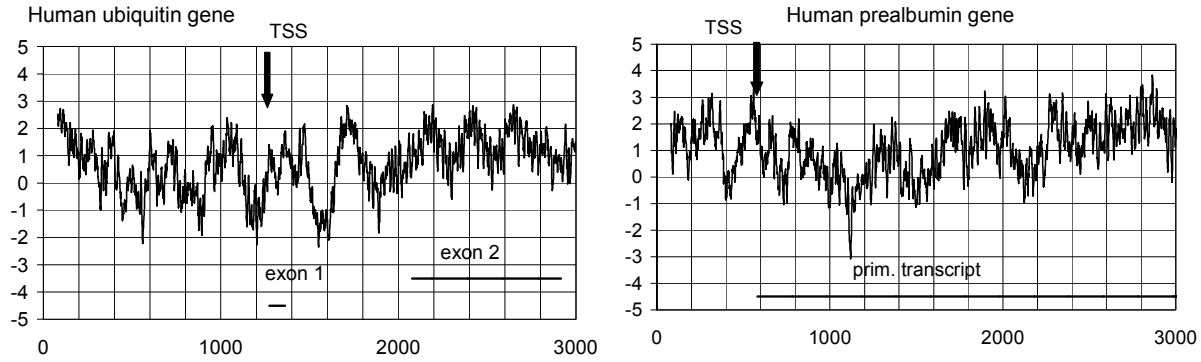
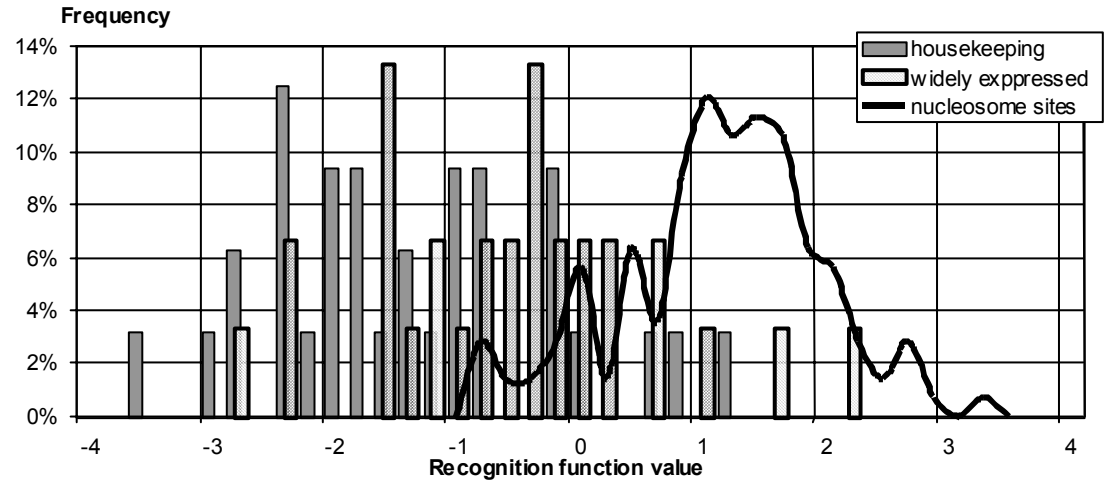


Figure 2. Nucleosome recognition function profiles for the ubiquitin (AC U49869) and prealbumin (AC M11844) genes.

a) Promoters of "housekeeping" genes and genes, which are widely expressed in various tissues



b) Promoters of tissue-specific genes

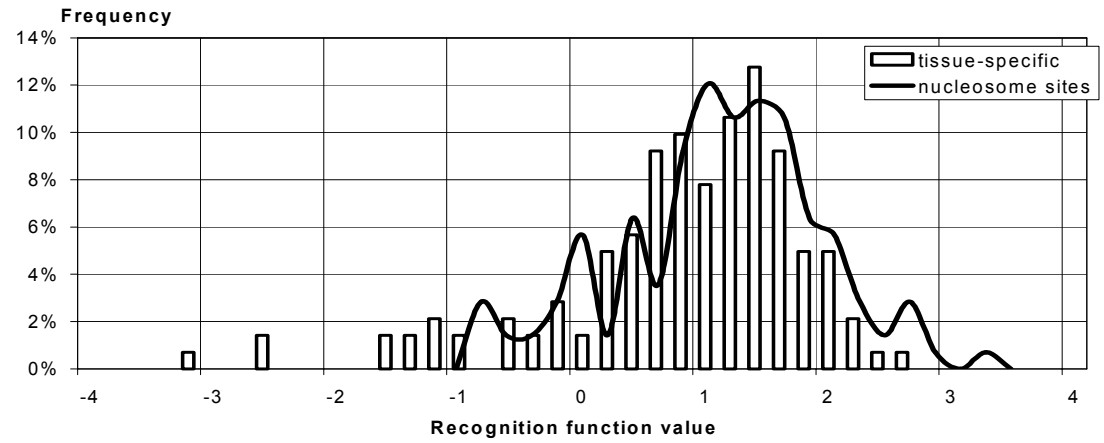


Figure 3. Histogram of distribution by the values of nucleosomal sites recognition function for promoter gene regions. For comparison, distribution of values of the same function for the training set of nucleosomal sites is given.

It could be noted that most of the promoters studied and extracted from the EPD database (Table 1), have the tissue-specific expression pattern. Possibly, this very phenomenon is related to the fact that periodicity of transcription factors binding, which is presumably explained by nucleosomal sites positioning, were recently found [Ioshikhes et al., 1999].

Thus, we have detected that among the factors regulating gene expression is ability of DNA to nucleosome positioning within transcription start regions. Moreover, gene expression pattern is likely governs in the course of evolution by such selection of nucleotide context within promoter region that is possible to provide nucleosome

packaging density optimal for gene functioning. For promoters of genes, expression of which should be finely tuned or limited (in case of tissue-specific promoters), selection in favor of tightly packed promoter regions takes place. In case it is not favorable to inhibit gene expression is some conditions (e.g., widely expressed or "housekeeping" genes), then nucleotide context is necessary, which should provide less dense nucleotide packaging or its complete absence.

Acknoledgments

This work was supported by the Russian Foundation for Basic Research (grants Nos 98-07-90126, 98-07-91078, 99-07-90203) and by the grant for Young Scientists by Siberian Branch of RAS.

References

1. Fitzgerald D.J., Dryden G.L., Bronson E.C., Williams J.S., Anderson J.N. (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. *J. Biol. Chem.,* 269, 21303-21314.
2. Ioshikhes I., Trifonov E.N. (1993) Nucleosomal DNA sequence database. *Nucl. Acids. Res.*, 21, 4857-4859.
3. Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., Trifonov E.N. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J Mol Biol.*, 262, 129-139.
4. Ioshikhes I., Trifonov E.N., Zhang M.Q. (1999) Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. *Proc Natl Acad Sci USA,* 96, 2891-2895.
5. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., Overton G.C. (2000) Transcription regulatory regions database (TRRD): its status in 2000. *Nucleic Acids Res.*, 28(1), 298-301.
6. Levitsky V.G., Ponomarenko M.P., Ponomarenko J.P., Frolov A.S., Kolchanov N.A. (1999) Nucleosomal DNA property database. *Bioinformatics,* 15 (7/8), 582-592.
7. Levitsky V.G., Katokhin A.V., Kolchanov N.A. (2000) Inherent modular promoter structure and its application for recognition tools development. *Computational technologies (Novosibirsk)*, 5, spec.issue, 41-47.
8. Paranjape S.M., Kamakaka R.T., Kadonaga J.T. (1994) Role of chromatin structure in the regulation of transcription by RNA polymerase II. *Annu Rev Biochem*, 63, 265-297.
9. Pedersen A.G., Baldi P., Chauvin Y., Brunak S. (1999) The biology of eukaryotic promoter prediction - a review. *Comput. Chem.*, 23, 191-207.
10. Perier R.C., Praz V., Junier T., Bonnard C., Bucher P. (2000) The eukaryotic promoter database. *Nucleic Acids Res,* 28, 302-303.
11. Satchwell S.C., Drew H.R. and Travers A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.,* 191, 659-675.
12. Sivolob A.V., Kharpunov S.N. (1995) Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *J. Mol. Biol,* 247, 918-931.
13. Stein A., Bina M. (1999) A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res.*, 27, 848-853.
14. Wolffe A.P. (1994) Nucleosome positioning and modification: chromatin structure that potentiate transcription. *Trends Genet.* 19, 240-244.

COMMON B-DNA FEATURES OF A DEFINITE TRANSCRIPTION FACTOR BINDING SITES SUPERCLASS

Ponomarenko J.V., Ponomarenko M.P.

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

Keywords*:* DNA, transcription factor, superclass, recognition

Resume

Motivation:

Transcription complex is a mosaic arranged through DNA-protein interactions and protein-protein ones, which are provided by transcription factors binding to overlapping binding sites (TF-sites). Since the sites differ by the context, the features of B-DNA may serve as candidates for TF-sites peculiarities responsible for their overlapping.

Results:

For transcription factor superclasses (i.e., Basic Domain, BD; Zinc-coordinating, Zn; Helix-turn-helix, HTH; and β-Scaffold, β-S) the specific B-DNA features of TF-sites were identified, and the tools were developed for their recognition.

Availability:

recognition of superclasses, http://wwwmgs.bionet.nsc.ru/mgs/programs/bdna/

Introduction

Transcription complexes are mosaics arranged through DNA-protein interactions and protein-protein ones, which are provided by transcription factors binding to overlapping binding sites (TF-sites) [1]. Since the sites differ by the context, the features of B-DNA may serve as candidates for TF-sites peculiarities, which provide their overlapping. In our previous paper [2], we detected correlation between two TF site dichotomies, "Zinc-coordinating/Helix-turn-helix" and "Low/High"-twisted DNA preferment. In the present work, we report about additional 11 pairs of such dichotomies: 3D-similar transcription factors bind to corresponding 3D-similar sites, this being in a good correspondence to experimental data on overlapping TF-sites, which are stored in databases TRANSFAC [3], TRRD [4], and COMPEL [5]. By the system B-DNA-Video, we detected B-DNA properties specific for TF-sites superclasses and developed the programs for their recognition.

Materials and Methods

We have studied 42 types of TF-sites (Table 1). The algorithm of analysis is given below. **Step 1:** DNA sequences were centered by experimentally detected transcription factor binding sites. **Step 2:** the sets of TF-sites referring to BD, Zn, HTH and βS superclasses were prepared. **Step 3:** the 50%-training sets were organized for TF-sites; for superclasses, they contain 12 sites of each type. **Step 4:** for each DNA $S=\{s_1...s_L\}$, we put the mean property $\mathbf{X_k}$ at DNA region [a; b] ($1\le a\le b\le L$; $1\le k\le 38$):

$$X_{kab}(S(i))=\Sigma_{i=a,b-1}\, X_k(s_i s_{i+1})/(b-a). \quad (1)$$

Each X_{kab} was tested for significance of discrimination of the training set of sites out of 1000 random DNA by 100 bootstrap-tests of 6 statistical criterions. If the p-th test of the q-th criterion is significant, then $\alpha<0.05$, X_{kab} was marked by the positive value $0<u_{pq}(X_{kab})\le 1$, otherwise, $-1\le u_{pq}(X_{kab})<0$; and then averaged:

$$U(X_{kab})=\{\Sigma_{p=1,100}\, \Sigma_{q=1,6}\, u_{pq}(X_{kab})\}/600. \quad (2)$$

Step 5: non-correlated properties $\{X_{kab}\}$ with the maximal $\{U(X_{kab})>0\}$ were selected. **Step 6:** we construct the function for TF-site (superclass) recognition within arbitrary DNA S, by the formula:

$$F(S(i))=\{\Sigma_{n=1,N}\, [X_n(S(i)) - a_n]/b_n\}/N \quad (3)$$

where: a_n, b_n are the heuristic coefficients normalizing the F(S) by its mean values for TF-sites and random DNA, which were equal to 1 and –1, correspondingly. Recognition rule is:

$$\text{IF } \{F(S_i)>0\} \text{ THEN } \{\text{in position i of the sequence S is the TF-site center }\}. \quad (4)$$

Step 7: we have tested the revealed properties and the recognition function on the control set. **Step 8:** If the results on the control set were significant, then the recognition function for TF-sites recognition is ready.

Table 1. Analyzed superclasses (N_T and N_C are the sizes of the training and control sets)

Code	TF-sites: number of sequences in the sets	N_T	N_C
BS	AP-1:74, ATF:28, c-Fos:21, c-Jun:22, C/EBP:108, CP-1: 51, CRE-BP1: 26, CREB:37, E2F: 9, MyoD: 16, NF-1: 99, NF-E2: 12, NF-IL6: 12, RF-X: 12, USF: 20	140	428
Zn	COUP: 14, ER: 22, GAGA: 7, GAL4: 15, GATA: 76, GR: 54, PR: 20, RAR: 15, RXR: 21, Sp1: 178, T3R: 21, YY1: 20	109	362
HTH	c-Myb:19, EN:12,ETS:14, HNF1:42, HNF3:10, HSF:7, IRF-1: 6, TTF-1: 7, OCT: 62	75	104
βS	E2: 20), MEF-2: 11, NF-kB: 36, SRF: 29, TCF-1: 6, TBP: 24	50	76

Results and Discussion

By treating 42 sets of TF-sites (Table 1) by the algorithm described above, we found B-DNA properties, the mean values of which for TF-sites were considerably higher (less) than for random DNA. In Table 2, 12 detected correlations between dichotomies of superclasses and "lesser/higher" values of B-DNA properties of TF-sites (compared to random DNA). To each TF-site type, there was assigned a vector $[d_{fm},]_{1\le m\le 12}$ of numbers $d_{fm}\in\{-1, 0, 1\}$, which mean that "the mean value of the m-th property of the f-th type of sites was less/equal/higher than that for random DNA". In Fig. 1, one can see the result of analysis of vectors $\{d_{fm}\}$ by the UPGMA method with the similarity measure " Euclidean distance", or dichotomy of sites. It significantly correlates to dichotomy of superclasses {BD+Zn}/{HTH+βS}, α<0,01. For six methods of claster analysis and five scores of similarity (the package STATISTICA), we have tested that all combinations of these methods produce TF-sites dichotomies correlating with the dichotomy {BD+Zn}/{HTH+βS}. It is of common knowledge that the proteins of superclasses BD and Zn bind DNA due to positive electrostatic charges at Arginine, Lysine, or Zinc, whereas the proteins of superclasses HTH and βS – by electro-neutral hydrogen bonds, van der Waals, and hydrophobic contacts. Thus, 3D-similar transcription factors bind 3D-similar TF-sites of DNA.

Analogously, we studied the sets of superclasses BD, Zn, HTH and βS. In Fig. 2, for each superclass a histogram is shown for the best (due to Formula 2) property of B-DNA of TF-sites. For the superclass BS (basic domain), the best such property was the probability to contact with the nucleosomal proteins, which are known to bind DNA by basic domains. For the superclass β-scaffold, the best property was the high Roll angle, corresponding to DNA bending, which is known from 3D-structures of DNA complexes with TBP, SRF and MEF-2 proteins of this superclass. By formulas (3) and (4) together with the revealed properties of B-DNA, we have developed the methods for the TF-sites superclass recognition. In Fig. 3, one can see the control results for their testing. As can be seen, the significant difference between histograms for the recognition function of BD superclass for TF-sites (light) and random DNA (dark), χ^2=1534.64; the control for superclasses Zn, HTH and βS is also significant (Fig. 3). The standard test "false positive recognition within 1000 bp of random DNA" (Table 4) resulted for superclasses the value of false positive estimate within the range from 5.73 to 8.98, which corresponds to the limits from 0.54 to 13.8 of the similar estimation for the known methods of TF-sites recognition [6, 7]. The standard test "under-prediction of the known sites" revealed the false negative estimate "1 out of 13" for the sites PU.1 (superclass HTH) and "5 out of 23" for the NFAT sites (superclass βS). In Fig. 4, there is an example of this test application for the sequence (EMBL: AC=X55987). Notably, the sites NFAT and PU.1 were not used for developing the methods of the superclass recognition (Table 1). This is the novelty of this work.

At the next step, we plan to apply suggested combinatorial variant for the site recognition, namely, recognition of biologically related sites, to analysis of tissue-specificity.

Table 2. Correlation of dichotomies of TF-sites superclasses and B-DNA properties.

Dicho-tomy	The values of properties for DNA sites in comparison to random DNAs			α
	X[#]	lesser («-»)	higher («+»)	
HTH / Zn	Ω	COUP, T3R, Sp1, ER, RAR, GR	OCT, IRF-1, EN, HNF1, HNF3	0.05
	h	GR, HNF3, HNF1, IRF-1, TTF-1, OCT	RXR, T3R, Sp1	0.05
	β	HSF, IRF-1, T3R, Sp1, GAGA, RXR, ER, GAL4, COUP, RAR	HNF3, OCT, EN, HNF1, GATA	0.05
	λ	GATA, HNF3, OCT, HNF1, IRF-1, TTF-1	YY1, COUP, RAR, T3R, RXR, ER, Sp1, c-Myb	0.05
	t	GATA, OCT,HNF1, IRF-1, HNF3, EN, MEF-2	ER, YY1, T3R, Sp1, RXR, GAGA	0.01
BD / HTH	γ	C/EBP,CP-1,NF-IL6,HNF3,IRF-1,HNF1,OCT	AP-1,RF-X,NF-1, CREB, USF, MyoD	0.05
	D	IRF-1, Sp1, MyoD, NF-E2, USF, CP-1, NF-1, AP-1, CREB, E2F, RF-X	OCT, HNF1, HNF3, C\EBP	0.05
	ω	IRF-1, HSF, EN, OCT, HNF1, E2F, C/EBP, CP-1	HNF3, ATF, USF, c-Fos, CREB, MyoD, AP-1, c-Jun, NF-E2, CRE-BP1	0.05
	F	E2F, CREB, USF, NF-1	NF-IL6, C/EBP, CP-1, OCT, HNF3, EN, IRF-1, TTF-1, HNF1	0.05
	S	C/EBP, HNF3, HNF1, OCT, IRF-1	NF-1, E2F, CP-1, MyoD, CREB, USF, CRE-BP1	0.05
Zn/BD	d	GATA, GAGA, Sp1, T3R, COUP, RXR, RAR	c-Jun, CREB, ATF, NF-IL6, USF	0.01
βS/hth	ρ	OCT, HNF3, HNF1, TTF-1, IRF-1	SRF, E2, TCF-1, NF-kB	0.01

[#]) Ω, twist; h, rise; β, bend angle; λ, persistent length; τ, melting temperature; s, d,γ and D - value, length, bendability of minor and depth of major grooves; ω, propeller; F, energy; ρ, roll.

Table 3. The number of false positives per 1000 bp of random DNA, N_{fp}/Kbp (by 100 tests).

Superclass	**Our method**	N_{fp}/Kbp		TF-site	**Traditional method**	N_{fp}/Kbp
BS	Formula (3)	7.53		NFATp/c	Weight matrix [6]	5.37
Zn	- « -	8.23		ADR1	MatInspector [7]	13.8
HTH	- « -	8.98		MATA1	- « -	0.54
βS	- « -	5.73		NIT1	- « -	2.53

Table 4. Recognition of TF-sites superclasses within the sequences (EMBL AC: start).

Site	Method	*Recognized*					Not recognized	
PU.1	HTH	12	D13263:359, M84477:384, X55987:475,	U63963:1738, X17463:644, S71481:1358,	M77675:836, Z25545:221, D26616:342,	M77875:862, U39637:930, M97811:310	1	AF03533:2999
NFAT	βS	18	V00536:61, X02910:494, M23442:1028, X14473:125, L07488:423,	X70058:533, X02910:517, U90652:436, X14473:151, X55987:406	L07488:181, X02910:531, U90652:480, X00695:1081,	X02910:450, M23442:931, U90652:582, U21135:194,	5	L07488:320, L07488:553, X03021:572, X03020:1088, X14473:194

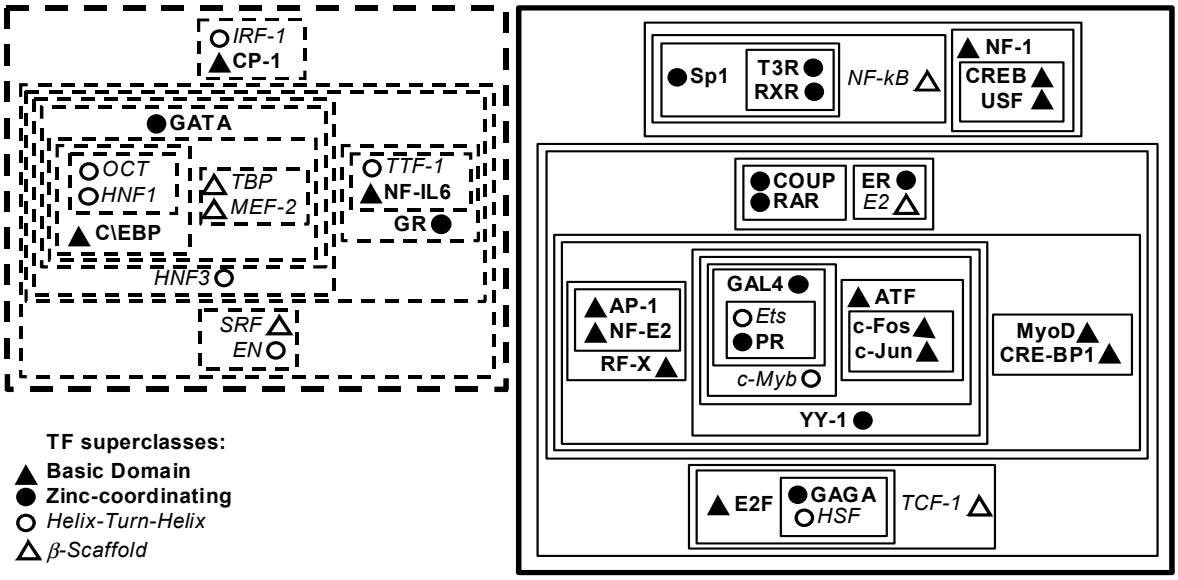


Figure 1. Cluster-analysis of 42 TF-sites. The method UPGMA. Similarity score "Euclidean distance.

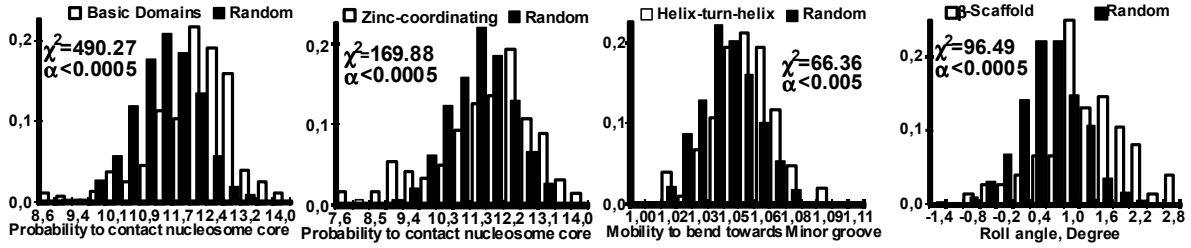


Figure 2. Histograms of the most significant B-DNA properties for TF-sites superclasses.

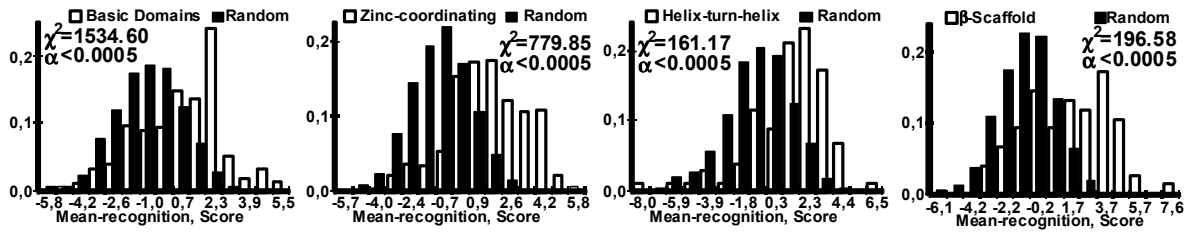


Figure 3. Histograms of mean recognition function of the superclasses of TF-sites.

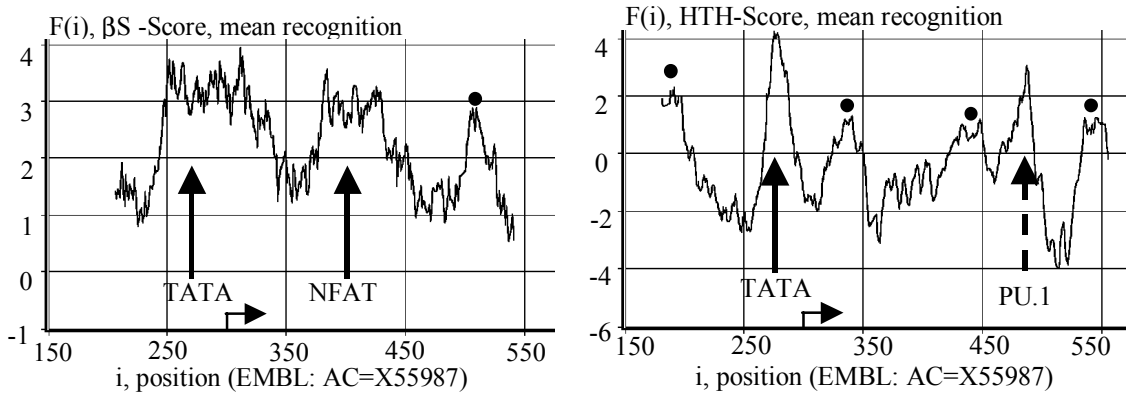


Figure 4. TF-sites recognition by the example of βS superclass (bold arrows: TATA-box and NFAT1) and HTH superclass (dotted arrow: PU.1). False positives are shown by dots.

The present study was supported by the Russian Foundation for Basic Research (grant No 98-07-90126).

References

1. P Johnson and S McKhight "Eukaryotic transcriptional regulatory proteins " *Annu. Rev. Biochem.* **58**, 799 (1989)
2. J Ponomarenko, M Ponomarenko, et al. "Conformational and physicochemical DNA features specific for transcription factor binding sites" *Bioinformatics.* **15**, 654. (1999)
3. E Wingender, X Chen, et al. "TRANSFAC: an integrated system for gene expression regulation" *Nucleic Acids Res*. **28**, 316. (2000)
4. N Kolchanov, O Podkolodnaya, et al. "Transcription regulatory regions database (TRRD): its status in 2000" " *Nucleic Acids Res*. **28**, 298. (2000)
5. O Kel-Margoulis, A Romashchenko, et al. "COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation" *Nucleic Acids Res*. **28**, 311. (2000)
6. A Kel, O Kel-Margoulis, et al. "Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells" *J. Mol. Biol.*, **288**, 353 (1999)
7. K Quandt, K Frech, et al. "MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data" *Nucleic Acids Res.,* **23**, 4878 (1995)

CONFORMATION OF TATA-PROMOTERS B-HELIX MAY GOVERN DIFFUSION OF TBP ALONG DNA TOWARDS -30 POSITION OF THESE PROMOTERS

Ponomarenko J.V., Pomonarenko M.P., Zvolsky I.L.

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

Keywords: promoter, TBP, TATA-less, non-specific binding, diffusion

Resume

Motivation:

As is known experimentally, specific TBP/TATA-box binding is preceded by the non-specific TBP binding with promoter and TBP diffusion along DNA. That is why an additional accounting of these properties may be useful for the TATA-box recognition.

Result:

A model of diffusion of TBP along DNA is constructed, where specific TBP/TATA-binding is a final state, whereas non-specific TBP/DNA-binding is a transitional state. Within the limits of this model, we have studied 891 promoters. It was shown that (1) TATA-box recognition score and TBP/DNA affinity have the peaks in -30 position of $TATA^{+}$ and $TATA^{-}$promoters; (2) the regions of promoters [-190; -45] and [-16; -3] have less TBP/DNA-affinity than random DNA; (3) TBP/DNA-affinity increases with the rank in evolution; (4) TATA-box recognition score decreases towards –30 position. For the yeast gene CYC1, it was shown that by taking into account additionally TBP/DNA diffusion, the number of false positive estimates in TATA-box recognition might be decreased.

Availability:

The TBP/DNA–affinity, http://wwwmgs.bionet.nsc.ru/mgs/programs/acts2/tbpdsdna.html. The TATA-box recognition score, http://wwwmgs.bionet.nsc.ru/mgs/programs/bdna/tata_bdna.html.

Introduction

As is known experimentally, specific TBP/TATA-binding is preceded by the non-specific TBP binding with promoter, this process being supplemented with TBP diffusion along promoter [1]. In the present paper, we have constructed a model of TBP/DNA diffusion, where specific TBP/TATA-binding is the final state, whereas non-specific TBP/DNA-binding is a transitional state. By applying this model, for 891 promoters it was shown that (1) TBP/DNA-affinity and TATA-box recognition score have the peak in position -30 of promoters $TATA^{+}$ and $TATA^{-}$; (2) TBP/DNA-affinity over the intervals [-190; -45] and [-16; -3] is less than in random DNA; (3) TBP/DNA-affinity grows with evolutionary rank; (4) recognition score of TATA-box diminishes to position -30. For promoter of the yCYC1 gene it was shown that taking into account diffusion of TBP decreases false prediction of TATA-boxes.

Materials and methods

We have studied 891 promoters (see Table), out of which after removing homologs [4], we made the sets of $TATA^{+}$ and $TATA^{-}$. The sequences were extracted according to description of promoters in [5]. Each DNA $S=\{s_1...s_L\}$ was characterized by the mean value $X_{kab}(S)$ of the property $X(s_is_{i+1})$ over the region [a; b], by the formula:

$$X_{ab}(S)=\Sigma_{1\le a\le i\le b-1\le L-1} X(s_is_{i+1})/(b-a). \quad (1)$$

To characterize DNA S, we also used the dinucleotide content $\phi\psi$, by the formula:

$$Y_{W\phi\psi ab}(S)= \Sigma_{1\le a\le i\le b-1\le L-1} W[(i-a)/(b-a)]\times\delta(s_i=\phi)\times\delta(s_{i+1}=\psi), \quad (2)$$

where: $0\le W[\pi]\le 1$ is a weight $\pi\in(0;1)$, $\delta(s=\xi)$ is a match-indicator: $\delta(\underline{true})\equiv 1$, $\delta(\underline{false})\equiv 0$.

Then, as was shown in our work [3], the regression TBP/DNA-affinity is as follows:

$$-\ln[K_D(S(i))] = -35.13 + 10.21\times X_{i-0,\, i+3}(S) - 0.72\times Y_{W,\, \phi=T,\, \psi=A,\, i-5,\, i+10}(S), \quad (3)$$

where i is a start of TATA-box, peak $W[\pi=0.75]=1$ is in the center (|) of the non-TATA 3'-tetranucleotide of TATA-box, tataAA|AA.

In our recent work [2], we have also optimized the TATA-box recognition score:

$$F(S(i))= [\Sigma_{1\le k\le 4} f_{k,a(i),b(i)}]/4, \quad (4)$$

where: $f_{k,a(i),b(i)} \equiv (X_{k,a(i),b(i)} - \gamma_k)/\tau_k$ is a partial recognition score of TATA-box by the B-DNA property; X; γ_k and τ_k are coefficients for normalizing f_k by means of ordering its mean values: "1" for TATA-boxes and "-1" for random DNA; $X_{1,i-8,i+4}$ – Bend of DNA [6], $X_{2,i-10,i+7}$ - tilt [6]; $X_{3,i-10,i+6}$ – Melting Temperature [7]; $X_{4,i-8,i+4}$ – Mobility to bend towards Minor groove [8].

Recognition rule: $\{F(S(i))>0\} \Rightarrow \{i - \underline{\text{start}}\ \underline{\text{T}}\text{ATA-box}\}$. (5)

The more detailed description of formulas (1-5) one can find in our recent works [2, 3].

By evaluating by formula (3) in position i of DNA S_n the TBP/DNA-affinity, $-\ln[K_D(S_n(i))]$ and after averaging it over the set N of $\{S_n\}_{1\le n\le N}$, we can obtain the profile of mean TBP/DNA-affinity, $-\ln[K_D(i)]$. Analogously, by the formula (4), we get the profile of mean recognition score for a TATA-box, F(i). The profiles F(i) and $-\ln[K_D(i)]$ corresponding to specific TBP/TATA-binding, were studied in our work [4]. In this work, we supplemented the formulas (3 and 4) by accounting of non-specific binding and TBP diffusion. In order to take diffusion into account, it is necessary to analyze the starting and transitional TBP/DNA-complexes [9], which in the simplest case correspond to specific TBP/TATA-binding and non-specific TBP/DNA with high affinity. As follows from this evidence, the lifetime of "transitional" complexes determine the time necessary for diffusion of TBP to the TATA-box. Let us denote $\Phi \in \{-\ln[K_D]; F\}$. Then the mean Φ_0 of the profile $\Phi(i)$ evidences that transitional complexes hindering TBP diffusion are absent; the upper boundary of deviations $\{\Phi_0 + t_{\alpha,\nu} \times s.d.\Phi(i)\}$ is the boundary of appearance of transitional complexes (here $t_{\alpha,\nu}$ is a Student coefficient; $s.d.\Phi(i)$ is a standard deviation in position i). Hence, formulas (3 and 4) may be improved by the simple accounting of TBP diffusion along DNA in a following way:

$$\Delta\Phi(i) = \Phi(i) - \Phi_0 - t_{\alpha,\nu} \times s.d.\Phi(i). \quad (6)$$

The novelty of the present work is the studying of $TATA^-$ promoters by the formula (6).

Implementation and Results

By the formula (6) for $TATA^-$ promoters, we have constructed the profiles of recognition score for the TATA-box, $\Delta F(i)$, and TBP/DNA-binding, $\Delta\{-\ln[K_D(i)]\}$ (Fig.1). In position -30 of $TATA^-$ promoters, we have found the peak indicating to the TBP-binding site contrary to the absence of the TATA-box. This is in a good agreement with experimental data (for review, see [10]). In this very peak, TBP/DNA-affinity corresponds to random DNA, within the regions [-190; -45] and [-16; -3], downstream the peak. This means that the peak of TBP/DNA-affinity in position -30 of $TATA^-$ promoters has appeared in the result of diminishing of this affinity in the regions indicated. Moreover, in these regions, the correlations were found between the TATA-box recognition score, $\Delta F(i)$, and i positions: to the left from position -30 this correlation was negative (r=-0.85), to the right – positive (r=0.89). Analogous tendency for the profiles $\Delta F(i)$ and $\Delta\{-\ln[K_D(i)]\}$ was observed for the $TATA^+$ promoters (see Table).

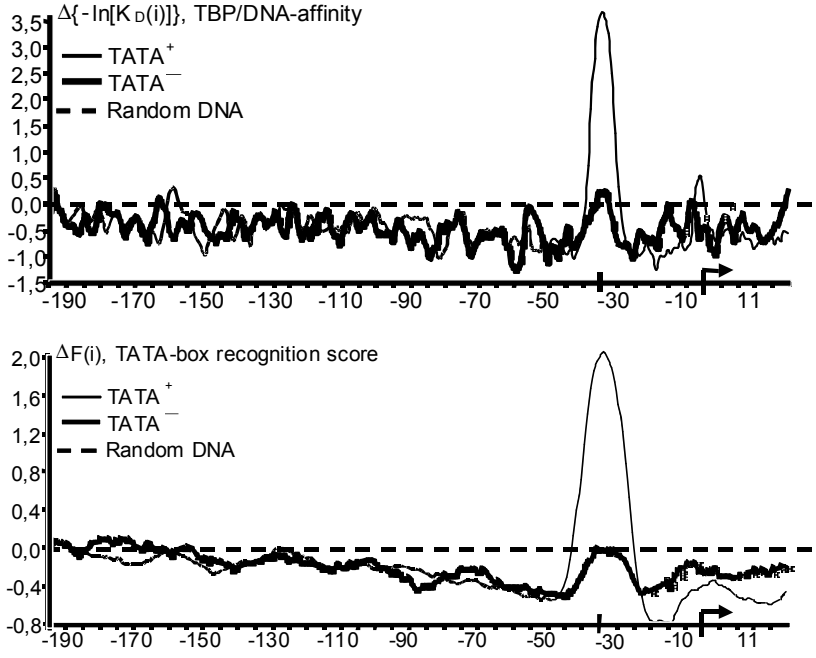


Figure 1. Profile of the TATA-box recognition score and TBP/DNA-affinity calculated by Formula (6).

Table. Parameters of TATA-box recognition score and TBP/DNA-affinity of promoters

Promoters		$\Delta[-\ln(K_D)]_0$, difference[$] from random DNA				Correlation of $\Delta F(i)$ and i			
Type	N	[-190; -45]		[-16; -3]		$i\in$ [-190;-45]		$i\in$ [-16; -3]	
		$\Delta[-\ln(K_D)]_0 \pm$s.d.	α	$\Delta[-\ln(K_D)]_0 \pm$s.d.	α	r	α	r	α
TATA$^-$	107	-0,46±0,23	$<10^{-25}$	-0,40±0,27	$<10^{-3}$	-0,92	$<10^{-3}$	0,96	$<10^{-3}$
Random DNA	30000	-0,03±0,02		0,03±0,02		-	-	-	-
Control:									
TATA$^+$	194	-0,41±0,24	$<10^{-22}$	-0,84±0,43	$<10^{-3}$	-0,85	$<10^{-3}$	0,89	$<10^{-3}$
Arthropodes	131	-0,46±0,42	$<10^{-12}$	-0,83±0,36	$<10^{-9}$	-0,76	$<10^{-3}$	0,98	$<10^{-3}$
Rest invertebrates[&]	45	-0,84±0,43	$<10^{-25}$	-1,32±0,47	$<10^{-5}$	-0,73	$<10^{-3}$	0,84	$<10^{-3}$
Birds	64	-0,68±0,35	$<10^{-23}$	-0,95±0,61	$<10^{-2}$	-0,86	$<10^{-3}$	0,92	$<10^{-3}$
Rodents	221	-0,37±0,21	$<10^{-23}$	-0,74±0,34	$<10^{-4}$	-0,90	$<10^{-3}$	0,94	$<10^{-3}$
Primates	205	-0,40±0,21	$<10^{-7}$	-0,59±0,32	$<10^{-3}$	-0,73	$<10^{-3}$	0,95	$<10^{-3}$
Rest mammals[@]	45	-0,76±0,51	$<10^{-22}$	-1,26±0,92	$<10^{-5}$	-0,48	$<10^{-3}$	0,96	$<10^{-3}$
Rest vertebrates[#]	33	-0,91±0,62	$<10^{-21}$	-1,16±0,80	$<10^{-2}$	-0,78	$<10^{-3}$	0,96	$<10^{-3}$
Plants	147	-0,40±0,26	$<10^{-21}$	-0,95±0,40	$<10^{-5}$	-0,70	$<10^{-3}$	0,88	$<10^{-3}$

N – volume of a set; [$]Student's criterion; i –number of position; $\Delta[-\ln(K_D)]_0$ and s.d. – mean and standard deviation; r –linear correlation coefficient; α - significance; [&]rest invertebrates (except artrapodes); [@]rest mammals (except rodents and primates); [#]rest vertebrates (except birds and mammals).

The presence of low affinity to TBP and correlations of TATA-box recognition score in vicinities of -30 position was tested on promoters of the following taxa: plants, arthropods, other invertebrates, birds, rodents, primates, other mammals, other vertebrates (see Table). As can be seen, promoters of all the taxa in the regions [-190; -45] and [-16; -3] have low TBP/DNA-affinity and decrease in the TATA-box recognition score towards –30 position. Thus, these two peculiarities of promoters detected over the set of TATA$^-$promoters were verified. Moreover, a relationships was found between mean TBP/DNA-affinity around position -30 and evolution of promoters (Fig. 2). Between the mean TBP/DNA-affinity within the regions [-190; -45] and [-16; -3] and evolutionary rank R of taxa: invertebrates (except arthropods) (R=1), birds (R=3), rodents (R=5), primates (R=6), other mammals (R=4), other vertebrates (R=2), - a linear correlations were found with r=0.88 and r=0.83, respectively (α<0.05). In detection of these correlations, we did not take into account the taxa of plants and arthropods, because the level of TBP/DNA-affinity there was intermediate between the levels of other invertebrates and rodents (Fig. 2).

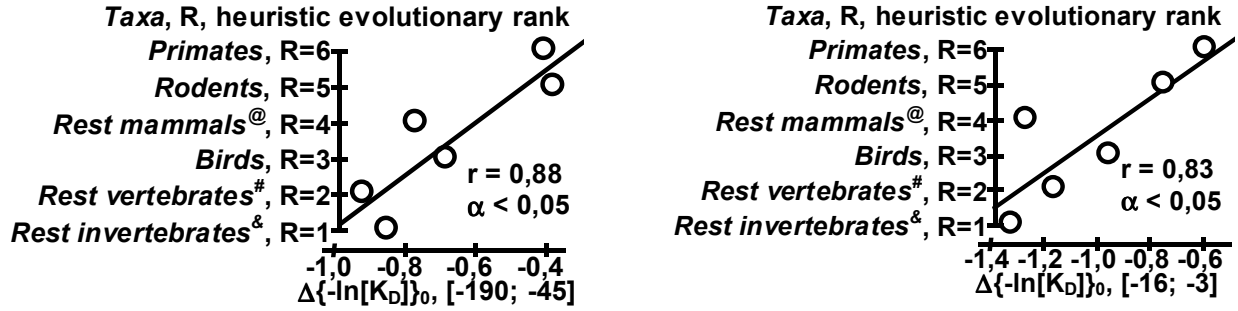


Figure 2. Linear correlation between the heuristic evolutionary rank and the TBP/DNA-affinity, $\Delta[-\ln(K_D)]_0$, calculated by Formula (6) and, then, averaged upon the TATA-less promoter regions, [-190; -45] and [-16; -3], flanking -30 position. Notes: r - linear correlation coefficient, α - significance; [&]rest invertebrates (except artrapodes); [@]rest mammals (except rodents and primates); [#]rest vertebrates (except birds and mammals).

As interpretation from the point of diffusion, the drift of TBP concentration towards this position may explain the decrease in TATA-box recognition score towards -30 position. With this respect, DNA conformation of TATA$^-$ promoters may direct diffusion of TBP towards -30 position. This is the novelty of the work presented.

Discussion and conclusion

By the example of yeast CYC1 gene promoter (EMBL: SCCYC1G5; X03472), by the formula (6), we have predicted the TATA-boxes with accounting of the non-specific binding and diffusion of TBP along promoter. The standard deviation for the formula (6) was calculated as:

$$s.d.F(i) \equiv [1/4 \times \Sigma_{1 \le k \le 4}\{F(S(i)) - f_{k,a(i),b(i)}(S)\}^2]^{1/2}, \quad (7)$$

where F(S(i)) is the TATA-box recognition score (by formula 4); $f_{k,a(i),b(i)}$ – partial TATA-box recognition score by the fixed property of B-DNA X_k (formula 4).

The result of formula (6) is shown in Fig.3 by bold line: all 5 occurring in nature TATA-boxes (circles) were recognized with a single false positive (arrow). This result obtained by the formula (6), we have compared to predictions made by the other methods recognizing TATA-boxes [2, 6, 11].

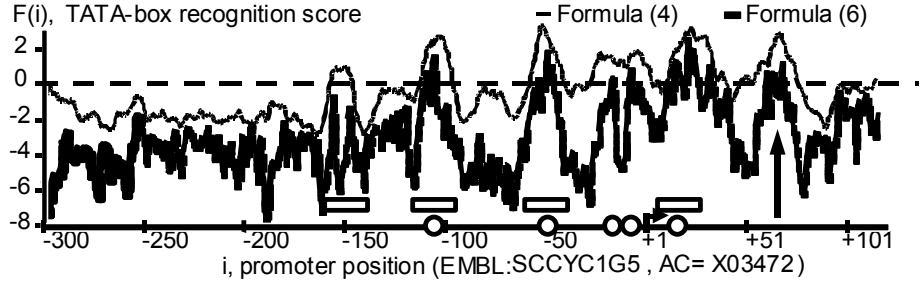


Figure 3. TATA-box recognition (circles) of promoter of the yeast CYC1 gene. The result by MatInspector method application [11] is marked by boxes. By arrow- the only false positive made by the formula (6).

In [6], TATA-boxes of promoter considered were recognized by heuristic search of regions with the wide minor groove. As a result, all natural TATA-boxes were recognized correctly with 28 false positives [6]. In [2], by formula (4), we have predicted TATA-boxes in this promoter (thin line): all 5 natural TATA-boxes were predicted, with 4 false positives. The method MatInspector [11], <http://transfac.gbf.de/cgi-bin/matSearch/matsearch.pl>, has predicted 13 TATA-boxes, which were located in 4 regions (Fig.3: boxes), where there were located 3 natural TATA-boxes (2 false negatives, 1 false positive). It is seen that by the formula (6), the natural TATA-boxes were predicted as by the methods [2] and [6], with 1 false positive as gave the MatInspector tools [11]. This means that adding of non-specific binding and diffusion of TBP along DNA (formula 6) leads to decrease in false positives.

In our future works we plan to use the formula (6) for recognition of the other DNA sites.

This work is supported by the Russian Foundation for Basic Research (No 98-07-90126). The authors are grateful to Galina Orlova for translation of the paper into English.

References

1. Coleman,R., and Pugh,B. (1995) Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA. J. Biol. Chem., 270, 13850-13859
2. Ponomarenko,J., et al. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics, 15, 654-668.
3. Ponomarenko,M. et al. (1999) Identification of sequence-dependent DNA features correlating to activity of DNA sites interacting with proteins. Bioinformatics, 15, 687-703.
4. Kolchanov N.A., et al. (1999) Integrated databases and computer systems for studying eukaryotic gene expression. Bioinformatics, 15, 669-686.
5. Perier,R., et al. (2000) The eukaryotic promoter database (EPD). *NAR,* 28, 302-303
6. Karas,H., et al. (1996) Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements. CABIOS. 12. 441-446.
7. Hogan,M.E., and Austin,R.H. (1987) Importance of DNA stiffness in protein-DNA binding specificity. Nature. 329. 263-266.
8. Gartenberg,M., and Crothers,D. (1988) DNA sequence determinants of CAP-induced bending and proteinRT binding affinity. Nature. 333. 824-829.
9. Eyring H. (1980) Basic chemical kinetics. John Wiley & Sons, New York. 528 p.
10. Smale,S. (1997) Transcription initiation from TATA-less promoters within eukaryotic promoter-coding genes. BBA, 1351, 73-88
11. Quandt,K., et al. (1995) MatInd and MatInspector - New fast and sensitive tools for detection of consensus matches in nucleotide sequence data. NAR, 23, 4878-4884.

THE MODULE ORGANIZATION OF THE A AND B BOXES IN THE tRNA INTRAGENIC PROMOTER

[1]Rogozin I.B., Kondrakhin Yu.V., Naykova T.M., Yudin N.S., [2]Voevoda M.I., *Romaschenko A.G.

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2]Institute of internal medicine SB RAMS, Novosibirsk, Russia
e-mail: romasch@bionet.nsc.ru
*Corresponding author

Keywords: tRNA genes, RNA polymerase III, promoter recognition

Resume

Motivation:

The eukaryotic tRNA gene promoters transcribed by RNA polymerase III consist of the two intragenic widely separated boxes A and B. Thus, the DNA gene composition of the A and B boxes regions share both the promoter and tRNA functions. Evolutionary transpositions of the promoter elements to the coding part of the tRNA genes must have brought about the appropriate evolving correction of promoter element structure that did not occur in prokaryotes because their promoters were restricted to the untranscribed regions. We compared the specific structural features of regions in sequences from various eukaryotes and prokaryotes. For this purpose, we developed a method for sequence alignment that not only took the conserved nucleotide into account, but also identified the most correlated positions.

Results:

Cluster analysis revealed two subclasses, A1 and A2, differing in deletion/insertion of a nucleotide at position 9 of the A box and also in nucleotide distribution at its other positions. It was shown that among eukaryotes the number of the 11 bp short box A variants considerably exceeds that of the 12 bp full-sized variants of the box. The relative proportions of the short to the full-sized variants were 61% and 31%, respectively. The homologous DNA sequences of the entire set of prokaryotic tRNA genes were in a proportion of 74% and 26%, respectively, in favor of the full-sized variants. Correlation analysis of nucleotide distribution at positions between and within the A and B boxes demonstrated that not only single positions, but also groups of adjacent positions, which we called modules, may be informational units for promoter recognition. This is reasonable since multiple correlations between the adjacent modules exist in eukaryotic tRNA genes, whereas the prokaryotic tRNA genes lack them. We statistically treated the distribution of a set polymorphic modules in the A and B boxes for the short A1 (11 bp) and full-sized A2 (12 bp) subclasses of the A box. Thus, to increase the recognition accuracy of the intragenic type 2 promoters of the RNA polymerase III transcribed genes, it is reasonable to take the A and B box module organization into account. This implies fuller consideration of the specific combinations of their modules in all the tRNA genes.

Introduction

Transcription functions in the genome have been distributed among RNA polymerases from the time of the eukaryotic nucleus emerged during evolution. The genes of the large ribosome subunits are transcribed by RNA polymerase I [1]. All proteins and the small nuclear RNAs, except the U6 RNA, are encoded by the RNA polymerase II transcribed genes [2]. 5S RNA, tRNA, U6 RNA, SINE and LINE repeats, as well as the adenovirus VAI genes, are transcribed by RNA polymerase III [3-9]; of these, type 1 and type 2 promoters are intragenic. Type 1 is characteristic of the 5S RNA gene, type 2 of the tRNA, SINE and VAI genes. Type 3 is present in the 5' flanking sequences of the U6 gene in most of the studied eukaryotes. Promoter elements have been additionally identified within the U6 gene in yeast [10].

The tRNA promoters consist of the two widely separated boxes A and B [11]. The A box is located on the stretch of DNA sequence which corresponds to a part of arm D of the spatial tRNA structure. The location of this part corresponds to the tRNA structure from the end of the acceptor stem, it includes the strand of the D duplex which becomes the D loop and three quarters of the D loop. The B box is located in the DNA region corresponding to the tRNA T-loop flanked at both ends by a nucleotide pair forming the tRNA T-stem [4]. The stretch between the A and B boxes contains the DNA region corresponding to the anticodon loop and the tRNA sequences flanking the regions at both ends, as well as the intron in the genes possessing it [12]. Only 10-20% of the tRNA genes have introns [13]. Thus, the DNA gene composition of the A and B boxes regions share both the promoter and tRNA functions. Evolutionary transpositions of the promoter elements to the coding part of the

tRNA genes must have brought about the appropriate evolving correction of tRNA gene structure that did not occur in prokaryotes because their promoters were restricted to the untranscribed regions.

Table 1. Distribution frequencies of module 1, 2 and 3 of the A box variants with the full-sized A1 and short A2 subclasses definite combinations in tRNA gene.

Variant			Module 2															Module 3											
			CCA	CCC	CGC	CGT	CTC	CTT	TCT	TGT	TTT	CCG	CGG	CCT	TAT	TCC	TTC	ATCGG	ATTGG	GATGG	GTAGG	GTCGG	GTTGG	GCTGG	ACGG	ATGG	GCGG	GGGG	GTGG
Module		TAG(105)	2	1	1	-	88	2	3	1	2	-						-	5	7	5	20	56	6					
		TGG(47)	5	4	16	3	1	2	4	8	1	-						6	2	2	-	10	20	1					
	Short	TAG(77)	-	1	-	-	22	2	1	20	5	-	-	-	17										3	5	6	7	53
		TGG(208)	-	6	38	3	6	-	33	22	2	62	-	7	4	8	7								5	47	37	9	106
		ATC(8)					8																						8
		TAA(4)					1	1			2															2		1	
		TCG(6)							3						3													1	5
		TGC(7)											7																7
		TGT(7)										6																	6
Module 2		CCA																					7						1
		CCC																					4			6			1
		CGC																5	2		3	3	4		2	11	14		11
		CGT																				2	1			3			
		CTC																1	2	6	2	20	50	6		5	1	7	23
		CTT																	1	1			1	1					2
		TCT																				2	3				1	9	30
		TGT																	1	1		4	4		3	13	9		17
		TTT																	1	1			1			6			3
		CCG																									18		49
		CGG																											8
		CCT																								7			
		TAT																										2	21
		TCC																											
		TTC																											7

We compared the specific structural features of the regions in sequences from various eukaryotes and prokaryotes. For this purpose, we developed a method for sequence alignment that not only took the conserved nucleotide into account, but also identified the most correlated positions [14]. Cluster analysis revealed two subclasses of the A box, A1 and A2, differing in deletion/insertion of a nucleotide at position 9 and also in distribution of nucleotides at its other positions. It was shown that the number of the 11 bp short box A variants considerably exceeds that of the 12 bp full-sized variants of the box. The relative proportions of the short to the full-sized variants were 61% and 31%, respectively. The homologous DNA sequences of the entire set of prokaryotic tRNA genes were in a proportion of 74% and 26%, respectively, in favor of the full-sized variants. Correlation analysis of nucleotide distribution at positions between and within the A and B boxes demonstrated that not only single positions, but also groups of adjacent positions, which we called modules, may be informational units for promoter recognition. This is reasonable since multiple correlations between the adjacent modules exist in eukaryotic tRNA genes, whereas the prokaryotic tRNA genes lack them.

Here we statistically treated the distribution of a set polymorphic modules in the A and B boxes for the A1 and A2 subclasses of the A box. We developed a criterion for the identification of the modules in the A and B boxes, the recognition of promoters and distinguishment of tRNA gene types on the basis of the identified modules. Examples are provided to show that the modules combine in a characteristic way with a particular type of the eukaryotic tRNA genes.

Methods and algorithms

DNA sequences from the databank available at http://www.unibayreuth.de/departments/biochemie/trna/ were used. The eukaryotic tRNA contains exhaustive information about the specific structural features of the A and B boxes in the tRNA genes. Additionally, the presence of tRNA in cell cytoplasm is direct evidence that the tRNA promoters are functionally active. Identification of tRNA gene type was based on transformation of the tRNA anticodon into DNA codon. The number of the eukaryotic A and B box samples was 645. The number of sequences homologous to the A and B boxes in the sample was 1034.

Based on correlation analysis, it was shown that the boxes can be divided into modules, with each containing a definite number of polymorphic variants. The frequencies of the occurrence and combination of various modules in the A and B boxes were treated statistically using the SPSS 8.0 program package.

Results and discussion

The high correlations among positions in the boxes and the distribution pattern of the invariant and relatively less conserved nucleotides allowed us to subdivide each of the A and B boxes into 3 modules. Highly conserved adenine at position 7 intervening between modules 2 and 3 was omitted. All the structural variants of modules 1, 2 and 3 of the B box combining with the polymorphic modules of the A box in the different subclasses are given in Tables 1 and 2. Clearly, module 1 is more diverse in the A2 (11 bp) than the A1 (12 bp) subclass with only 2 variants -TAG- and -TGG- , with -TAG- being twice more frequent in A1 subclass.

Remarkably, the -TAG- and -TGG- frequencies are reverse in the A1 subclass. In prokaryotes, the structural variations of module 1 are wider and in addition to the above variants, 8 other occur. In contrast to eukaryotes, they combine with long module 3 variants with an insertion at position 9 in the A box.

Table 2. Distribution frequencies of module 1, 2 and 3 of the B box variants with the full-sized A1 and short A2 subclasses definite combinations in tRNA gene.

	Variants of the B box	Module 2 of the B box							Module 3 of the B box									
		AAA	AAG	AAT	GAA	GAC	GAG	GAT	ACC	CCC	CCT	TCA	TCC	TCT	TCG	GCC	CCG	CCA
Full-sizes A1 subclass, Module 1	AGTTC(24)	2			4		7	10			16			8				
	GTTC(3)				1	2									3			
	GATC(6)						1	5	4							2		
	GTTC(188)	5	9	8	17	3	29	37	1	58			47	1				
	TTTC(11)							11				1						10
Full-sizes A1 subclass, Module 2	AAA(7)												5	2				
	AAG(9)									1			8					
	AAT(8)									5			3					
	GAA(22)									3	2		12	3	1			
	GAC(3)									1			2		2			
	GAG(39)								1	27	4		2	3		1		
	GAT(64)								5	21	9		15	1		2		10
Short-sized A2 subclass, Module 1	AGTTC(44)	14		9	13		1	7			11		1	32				
	GGTTC(2)				1			1					1				1	
	GGATC(28)	1			25			1	27	1								
	GGTTC(237)	23	2	34	92	17	14	54	6	44			181	4				
	TGTTC(7)				5			2				5						1
Short-sized A2 subclass, Module 2	AAA(38)								1	3			20	14				
	AAG(2)												2					
	AAT(40)									8	1		26	8				
	GAA(148)								52	16	6	5	70	8				1
	GAC(17)									4			12	1				
	GAG(14)									3			11	1				
	GAT(65)									11	4	2	41	6			6	

The long variants of the A box have less heterogeneous module 2. The most abundant variant are -CTC- and -CGC-. Much higher heterogeneity was found for module 2 of subclass 2. Structural variants -CCG-, -TGT-, -CGC-, -TCT-, -CTC-, -TAT- occur frequently (Tab. 1). As for prokaryotes, 22 structural triplet variants additionally occur in module 2.

Module 3 in subclass 1 is represented mainly by -GTTGG- and -GTCGG- variants. Variant -GTGG- predominates in the A2 subclass, variants -ATGG- and -GCGG- are smaller in number. In prokaryotes, additional 39 full-sized polymorphic variants were detected.

Thus, we showed that the structural diversity of the A box sequences in prokaryotes by far surpasses that in eukaryotes. The smaller structural diversity and the restricted combinations of the polymorphic A and B boxes are presumably the consequences of the promoter functions of the eukaryotic tRNA genes.

The DNA sequences in the B box region are much more conserved than those in the A box both in prokaryotes and eukaryotes. Applying the same module criterion, the box B was subdivided into 3 modules. Module 1 in the tRNA genes is present in all eukaryotic kingdoms, mainly as 5 polymorphic variants (Tab. 2). Three other structural variants were rarely detected in unicellular organisms and plants. The B box is also highly conserved in prokaryotes in which 4 pentamers were additionally detected in few numbers. Module 2 of the B box is the most polymorphic. It shows 7 structural variants, most of which are highly abundant in the eukaryotic B box. Only 4 variants were additionally detected in prokaryotic module 2.

The distribution frequencies of the module 2 structural variants depends on the A box subclasses with which the B box combines. In the case of the full-sized A boxes, the structural variants of the B box module 2, -GAT- (42%), -GAG-(24%) and -GAA- (14%) were most abundant. When the box A was referred to the A2 subclass, the frequencies of module 2 variants were -GAA- (44%), -GAT- (20%), -AAT- (13.5%), -AAA- (12%), -GAG- (5%).

Correlations were established also between the box A subclasses (the A1 and A2) and the box B module 3. Of the 10 structural variants of module 3, -CCC- and –TCC- were predominant in the B box combination with the A1 and A2 subclasses. However, when the B box combined with the full-sized A1, 38% of the B box samples contained –CCC-, when it combined with the short A2, only 13% contained –CCC-. The –TCC- variant predominated (59%) in the B box with short A2 box combination and its frequency was 31% in the B box with

full-sized A1 combination. The occurrence frequencies of the box B module 3 another variants varied, depending on the box A subclasses it combined with. The percentage for the combination were as follows: the –ACC- with A1 subclass, 3%, -ACC- with A2 subclass, 10,3%; the –CCT- with A1 subclass, 10.7%; -CCT- with A2 subclass, 3%; the –TCT- with A1 subclass, 6%, -TCT- with A2 subclass, 11.3%; the –CCA- with A1 subclass, 6.5%, -CCA- with A2 subclass, 0.3%. In all, we identified 6 structural variants of module 3 at high occurrence frequencies in the B box, which vary depending on the A subclasses.

The considerable decrease in the occurrence frequencies of the full-sized A variant (the A1) subclass and increase in those of the short ones (the A2) subclass in the eukaryotic tRNA gene promoters could be hardly attributable to modified translational activity of tRNA. The changes rather resulted from a promoter newly arisen on this DNA sequence. This explanation appears reasonable in view of the decreasing diversity of the A box variants during the evolutionary advancements from prokaryotes to eukaryotes. This prompted us to see how a small number of the full-sized A box variants may be distributed among the tRNA genes of various types. It proved that the sequences of the various types could be assigned to 3 groups according to whether the full-sized variants were present or not in the promoter. Group I, only full-sized box A variants: tRNALys(AAG),(AAA); tRNAPhe(TTC); tRNATyr(TAC); tRNAIle(ATT),(ATA) tRNA genes. The tRNA genes of single cell and animal kindoms all possess Group I. Exceptions are the tRNALys(AAA), tRNATyr(TAC) and tRNAIle(ATT),(ATA) genes because the short A box variants occur in these genes, very occasionally, in unicellular organisms.

Group II, the genes whose promoters contain only the short A box variants. These are the tRNASer(AGC),(TCA); tRNAVal(GTG),(GTA); tRNAPro(CCA),(CCT); tRNAAsp(GAC); tRNAGln(CAG),(CAA); tRNAGly(GGC); tRNAHis(CAC); tRNALeu(TTG),(CTG),(CTT),(TTA); tRNAArg(AGA),(CGT) genes. The full-sized A box variants were found to occur infrequently in the tRNAArg(AGA),(CGT); tRNALeu(TTA); tRNAVal(GTA) and tRNAGln(CAA) genes in unicellular organisms.

Group III, the genes containing both the full-sized and short variants of the A box. These include the tRNALeu(GCT),(GCA); tRNAThr(ACT); tRNAAsn(AAC); tRNAVal(GTT); tRNAMet(ATG); tRNAGlu(GAA),(GAG) and tRNATrp(TGG) genes. The data on the tRNA genes other than those noted above were not used because of the small sample size in the database.

The specificities of combination variants of the different A and B box modules depend on the A box subclass (Table 1,2). For example, the tRNALys genes for both AAA and AAG codons have just the full-sized variant in the promoter. The presence of the structural variant –GTCGG- in the box A module 3 is a characteristic feature of these genes. –GTCGG- is specific to the tRNA promoters of all the, so far, analyzed multicellular organisms. Thirty sequence variants with –GTCGG- in module 3 occurring in the tRNA gene samples are set out in Table 1. Of the 30, 24 were identified in the tRNALys genes, and 6 were detected in the tRNAAsn , tRNAGlu, tRNAAla, tRNAVal and tRNATre genes of unicellular organisms. The 6 were found to be combined with the structural variants of module 1 –TGG- and also with 4 variants of box A module 2 (-CGC-, -TGT-, -CGT-, -TCT-). Modules 1 and 2 in the tRNALys genes occur only as –TAG- and –CTC- variants, respectively, that predominate only in the full-sized A box variants. The sequence –TAGCTCAGTCGG- of the A box is conserved in the tRNALys genes of multicellular organisms.

The two types of the tRNALys(AAA) and (AAG) genes differ in the structural variants of the B box modules 2 and 3. Polymorphic variants of the B box modules 2 –GAG- and 3 –CCC- are present in the tRNALys(AAG) gene. In contrast to the tRNALys(AAG) gene, the other tRNALys(AAA) gene occurs in combination with –AAG- and –TTC-, respectively (see Table 2). Hence, the structure of the B box in the tRNALys(AAG) gene is –GGTTCGAGCCC- and the one in tRNALys(AAA) gene is –GGTTCAAGTCC-. Remarkably, evolutionary distant multicellular organisms share this structure.

The conservation of particular combinations of the A and B box module variants is an obligatory structural feature of the intragenic tRNA gene promoters in multicellular organisms. The structure of the A and B boxes in the tRNAMet(ATG) gene has mostly the following structure in multicellular organisms. When the tRNAMet(ATG) gene promoter contains the short A box variant –TGGCGCAGCGG-, it surely contains the B box with the structure –GGATCGAAACC-. In the case when it contains the full-sized A box –TAGCGCAGTAGG-, the structure of the B box is –AGTTCGATCCT-. The box A module 3 variant –GTAGG- is listed five times in Table 1. All the 5 variants are present in the A box of the eukaryotic tRNAMet genes. Combinations of the A and B box modules are specific to each of the analyzed tRNA gene types.

Thus, to increase the recognition accuracy of the intragenic type 2 promoters of the RNA polymerase III transcribed genes, it is reasonable to take the A and B box module organization into account. This implies fuller consideration of the specific combinations of their modules in all the tRNA genes.

References

1. Sentenac A. (1985) Eukaryotic RNA polymerases. CRC Crit. Rev. Biochem, 18, 31-91.
2. Zawel L., Reinberg D. (1992) Advances in RNA polymerase III transcription. Curr.Opin. Cell Biol, 4, 488-495.

3. Setzer D.R., Brown D.D. (1985) Formation and stability of the 5' RNA transcription complex. J. Biol. Chem,, 260, 2483-2492.
4. Sharp S., DeFranco D., Dingermann T., Farrell P., Soll D. (1981) Internal control regions for transcription of eukaryotic tRNA genes. Proc. Natl. Acad. Sci. USA, 78, 6657-6667.
5. Eschenlauer J.B., Kaiser M.W., Gerlach V.L., Brow D.A. (1993) Architecture of a yeast U6 RNA gene promoter. Mol. Cell Biol, 13, 3015-3026.
6. Geiduschek P.E., Tocchini-Valentini G.P. (1988) Transcription by RNA polymerase III. Ann. Rev. Biochem., 57, 873-914.
7. Kurose K., Hata K., Hattori M., Sakiki Y. (1995) RNA polymerase III dependence of the human L1 promoter and possible participation of the pol II factor YY1 in the RNA polymerase III transcription system. Nucl. Acids Res, 23, 3704-3709.
8. Lobo S.M., Tanaka M.L., Sullivan M.L., Hernandez N. (1992) A TBP complex essential for transcription from TATA-less but not TATA-containing RNA polymerase III promoters is part of the TFIIIB fraction. Cell, 71, 1029-1040.
9. Gabrielsen O.S., Sentenac A. (1991) RNA polymerase III(C) and its transcription factors. TIB, 16, 412-416.
10. Brow D.A., Guthric C., (1990) Transcription of a yeast U6 snRNA gene requires a polymerase III promoter element in a novel position. Genes Dev, 4, 1345-1356.
11. Rich A., Rajbhandary U.L. (1976) Transfer RNA: molecular structure, sequence and properties. Ann. Rev. Biochem, 45, 805-860.
12. Sharp S.J., Schaack S., Cooley L., Burke D.S., Soll D. (1985) Structure and transcription of eukaryotic tRNA genes. . CRC Crit. Rev. Biochem, 19, 107-144.
13. Guthrie C., Abelson S. (1982) Organization and expression of tRNA genes in *Saccharomyces cerevisiae*. In "The Molecular Biology of the Yeas *Saccharomyces*, Metabolism and Gene Expression, Cold Spring Harbor Laboratory, New York, 487.
14. Kondrakhin Yu.V., Rogozin I.B., Romaschenko A.G. This issue.

B-DNA FEATURES CORRELATING WITH POINT MUTATIONS THAT INFLUENCE DNA/PROTEIN-BINDING FREE ENERGY

Ponomarenko M.P., Ponomarenko J.V., Goryachkovskaya T.N., Orlova G.V., [1]Sarai A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: pon@bionet.nsc.ru
[1]Institute of Physico-Chemical Research (RIKEN), Japan
*Corresponding author

Keywords: DNA/protein-binding, energy, nucleotide polymorphism, BDNA, single nucleotide polymorphism

Resume

Motivation:

Databases accumulating information on single nucleotide polymorphism detected by alterations in gene expression make timely an analysis of point mutations in DNA/protein-complexes.

Results:

The data were studied on the point mutation impact on the energy ΔΔG of DNA binding with Cro- and λ-repressors, INR, c-Myb, EREBP-2, and AtERF-1. It was shown that substitutions influencing on binding affinity and specificity differ by localization and ΔΔG-patterns. The correlations between ΔΔG of point mutations and BDNA features were found. These correlations make the grounds for developing of the methods aimed at ΔΔG prediction for multiple substitutions and at recognition of DNA/protein binding sites.

Availability:

ACTIVITY database, URL=<http://wwwmgs.bionet.nsc.ru/systems/activity/>.

Introduction

Development of databases accumulating the knowledge on single nucleotide polymorphism detected by gene expression disturbances [1, 2] makes topical the studying of point mutations within the complexes DNA/protein. For substitutions G663A and G666T located in intron 6 of the *TDO2* gene and causing mental disorders, it was demonstrated that they damage YY1-site, as was proved by experiments in studying affinity of modified DNA to the nuclear extract from rat liver and by experiments using specific antibodies [3]. With this respect, it would be appropriate to analyze the information, accumulated in our database ACTIVITY [4], that concerns point mutation impact on the energy ΔΔG of binding DNA with Cro- and λ-repressors [5, 6], proteins INR [7], c-Myb [8, 9], EREBP-2 and AtERF-1 [10].

In the present paper, we have demonstrated that the substitutions influencing the affinity (50% with the lesser ΔΔG) and specificity (50% with the higher ΔΔG) differ both by localization within the sites and by ΔΔG-patterns in various test-systems. The correlations were found between ΔΔG of point mutations and conformational and physico-chemical DNA properties [11-16]. They are laid in foundation of the methods developed for prediction of ΔΔG for multiple substitutions and recognition of DNA-sites.

Materials and Methods

All the data analyzed are shown in Table 1 (the format of ΔΔG-matrices is given in [17]). The set of sites with the known DNA sequences and values of DNA/protein energy, $[S_n=\{s_{n,i}\}, \Delta\Delta G_n]$, was introduced into the system ACTIVITY [4], which performs an automated search for the most justified simple correlation:

$$\Delta\Delta G_n = F_0 + F_1 \times [\Sigma_{a \le i \le b-1} X(s_{i,n} s_{i+1,n})/(b-a+1)], \quad (1)$$

where [a;b] is a site fragment; X, one of BDNA features accumulated in ACTIVITY [4].

Following the theory of decision making [18], for the measure of validity was taken the ratio of statistical criterions, of applicability of a simple correlation [19], which are satisfied on the set $[S_n, \Delta\Delta G_n]$.

Results and Discussion

By dichotomy, we have subdivided all point mutations into two types (Table1): by *italic* are marked 50% of *lesser* ΔΔG (they may influence *affinity*), **in bold** are given 50% of **higher** ΔΔG (they influence **specificity**). We have found the discrepancies between substitutions of these types by their localization and ΔΔG-patterns (Table 2). Due to these discrepancies, if ΔΔG of point mutation that influences an affinity correlates to some DNA

property, then $\Delta\Delta G$ of substitutions influencing specificity is most likely not to correlate with this property. Hence, in our work, the point mutations influencing affinity and specificity were studied independently. The search for relationships "DNA$\Rightarrow\Delta\Delta G$" for point mutations influencing affinity separately from the same relationships for mutations modifying specificity is a novelty of the approach suggested. The results of Cro- and λ-repressors, INR, c-Myb, EREBP-2 and AtERF-1 analyses are given in Figure 1.

For the point mutations in the O_R1/Cro-complex [5], the most justified were the correlations:

$\Delta\Delta G_A=11.40+0.87\times[\Sigma_{-10\le i\le 9}\ \omega(s_is_{i+1})/20]$, IF substitution influences affinity (Fig.1a), (2)

$\Delta\Delta G_S=-2.95+0.46\times[\Sigma_{-7\le i\le 9}\ P(s_is_{i+1})/18]$, IF substitution influences specificity (Fig.1b),

where: DNA, 5'-TT[O_R1]AT-3'; ω, P, BDNA features (Table 3); Fig.1c, conjunction $\Delta\Delta G_A\oplus\ \Delta\Delta G_S$.

Additional data on $\Delta\Delta G_{Mult}$ of multiple substitutions [5] lead to regression given in (Fig.1d):

$\Delta\Delta G_{MULT}=-0.03+1.05\times Score_{\Delta\Delta G}-\delta\times[0.05\times MIN\{\Delta\Delta G_{50\%}:\ \Delta\Delta G_A\}+0.10\times MAX\{\Delta\Delta G_{50\%}:\Delta\Delta G_S\}]$, (3)

where: $Score_{\Delta\Delta G}$, the sum of $\Delta\Delta G$-matrix for DNA; $\Delta\Delta G_{50\%}$, median; δ, share of non-coincidence.

For the point mutations in O_R1/λ-complex [6], the most justified were the correlations:

$\Delta\Delta G_A=-6.98-0.60\times[\Sigma_{-9\le i\le 9}\ \omega(s_is_{i+1})/19]$, IF susbstitution influences affinity (Fid.1d), (4)

$\Delta\Delta G_S=-10.65+4.18\times[\Sigma_{-8\le i\le 7}\ \beta(s_is_{i+1})/16]$, IF susbstitution influences specificity (Fig.1e),

where: $\Delta\Delta G_A\oplus\ \Delta\Delta G_S$, Fig.1g; regression $\Delta\Delta G_{Mult}$ of multiple substitutions [6] was as in (Fig.1h):

$\Delta\Delta G_{MULT}=0.50\times Score_{\Delta\Delta G}(S)+\delta_{\#}\times[0.25\times MIN\{\Delta\Delta G_{50\%}:\ \Delta\Delta G_A\}+0.25\times MAX\{\Delta\Delta G_{50\%}:\Delta\Delta G_S\}]$. (5)

For the point mutations in Inr-site [7], the best simple correlations looked as:

$\Delta\Delta G_A=-0.33+0.004\times[\Sigma_{-3\le i\le 3}\ D(s_is_{i+1})/6]$, IF susbstitution influences affinity (Fig.1i), (6)

$\Delta\Delta G_S=-69.42+2.01\times[\Sigma_{-2\le i\le 4}\ \Omega(s_is_{i+1})/6]$, IF susbstitution influences specificity (Fig.1j),

where:$\Delta\Delta G_A\oplus\Delta\Delta G_S$, Fig.1k; by substitution (6$\Rightarrow$5), $\Delta\Delta G_{MULT}$ is well predicted from [7] (Fig.1l).

For the point mutations in c-Myb [8, 9], the best simple correlations looked as:

$\Delta\Delta G_A=23.69-21.81\times[\Sigma_{1\le i\le 4}\ \mu(s_is_{i+1})/4]$, IF susbstitution influences affinity (Fig.1m), (7)

$\Delta\Delta G_S=3.44+0.59\times[\Sigma_{1\le i\le 3}\ \tau(s_is_{i+1})/3]$, IF susbstitution influences specificity (Fig.1n),

where:$\Delta\Delta G_A\oplus\Delta\Delta G_S$, Fig.1o; (7$\Rightarrow$5) enabled to recognize c-Myb-sites from [20, 21] (Fig.1p).

For the point mutations in the plant GCC-box [10], the best simple correlations looked as:

$\Delta\Delta G_A=-69.94+70.20\times[\Sigma_{-5\le i\le 1}\ \tau(s_is_{i+1})/6]$, IF susbstitution influences affinity (Fig.1q), (8)

$\Delta\Delta G_S=-5.67-0.99\times[\Sigma_{-3\le i\le 3}\ \omega(s_is_{i+1})/6]$, IF susbstitution influences specificity (Fig.1r),

where:$\Delta\Delta G_A\oplus\Delta\Delta G_S$, Fig.1s;DNA, 5'-G[GCC-box]A-3', (8$\Rightarrow$5) recognized GCC-boxes (Fig.1t).

For the point mutations influencing affinity and specificity, $\Delta\Delta G$ correlates with different DNA properties in all the cases. The tests on the control (Fig.1: a,b,e,f) and undependable (Fig.1: i,j,m,n,q,r) data, predictions of multiple substitutions (Fig.1l) and site recognition (Fig.1: p,t) demonstrate the significance of relationships "DNA$\Rightarrow\Delta\Delta G$" for point mutations.

The present study was supported by Russian Foundation for Basic Research (98-07-90126) and STA Fellow #499042 (Japan).

References

1. Smigielski E.M., Sirotkin K., Ward M., Sherry S.T. (2000) *Nucleic Acids Res.,* 28, 352-355.
2. Brookes A.J., Lehvaslaiho H., Siegfried M., et al. (2000) *Nucleic Acids Res,* 28, 356-360.
3. Vasiliev G.V., Merkulov, V.M., Kobzev V.F., et al. (1999) *FEBS Lett.,* 462, 85-88.
4. Ponomarenko MP., Ponomarenko JV, Frolov A.S., et al. (1999) *Bioinformatics,* 15, 687-703.
5. Takeda Y., Sarai A., Rivera V.M. (1989) *Proc. Natl. Acad. Sci. USA*., 86, 439-443.
6. Sarai A., Takeda Y. (1989) *Proc. Natl. Acad. Sci. USA*., 86, 6513-6517.
7. Kraus R.J., Murray E.E., Wiley S.R., et al. (1996) *Nucleic Acids Res.,* 24, 1531-1539.
8. Tanikawa J., Yasukawa T., et al. (1993) *Proc. Natl. Acad. Sci. USA.,* 90, 9320-9324.

9. Ogata K., Kanei-Ishii C., Sasaki M., et al. (1996) *Nat. Struct. Biol*., 3. 178-187.
10. Hao D., Ohme-Takagi M., Sarai A. (1998) *J. Biol. Chem*., 273, 26857-26861.
11. Karas H., Knuppel R., Schulz W., et al. (1996) *Comput. Appl. Biosci*., 12, 441-446.
12. Gorin A.A., Zhurkin V.B., Olson W.K. (1995) *J. Mol. Biol*., 247, 34-48.
13. Suzuki,M., Yagi,N., Finch J. (1996) *FEBS Lett*. 397, 148-152.
14. Shpigelman E.S., Trifonov E.N., Bolshoy A. (1993) *Comput. Appl. Biosci*., 9, 435-440.
15. Satchwell,S.C. Travers,A.A. (1989) *EMBO J*., 8, 229-238.
16. Gartenberg M.R., Crothers D.M. (1988) *Nature*, 333, 824-829.
17. Deng Q.L., Ishii S., Sarai A. (1996) *Nucleic Acids Res.*, 24, 766-774.
18. Fishburn P.C. (1970) *Utility theory for decision making*, New York: Jonh Wiley & Sons.
19. Lehman E.L. (1959) *Testing statistical hypotheses*. New York: John Wiley & Sons.
20. Kolchanov N.A., Podkolodnaya O.A., et al. (2000) *Nucleic Acids Res.*, 28, 298-301.
21. Ponomarenko J.V., Orlova G.V., et al. (2000) *Nucleic Acids Res.,* 28, 205-208.

Table 1. Data analyzed on the influence of point mutations on the energy ΔΔG of binding DNA-protein

Binding of the operator O_R1 with the Cro-repressor [5]

s_i	T_{-8}	A_{-7}	C_{-6}	C_{-5}	T_{-4}	C_{-3}	T_{-2}	G_{-1}	G_0	C_{+1}	G_{+2}	G_{+3}	T_{+4}	G_{+5}	A_{+6}	T_{+7}	A_{+8}	A_{+9}
A	*0.12*	0	*-0.74*	**1.97**	*-0.59*	**1.90**	**1.95**	*-0.18*	*0.59*	*-0.24*	*0.39*	**2.30**	**1.20**	**1.37**	0	**2.50**	0	0
T	0	**1.85**	*-1.30*	**1.68**	0	**2.28**	0	*-0.32*	*0.64*	*-0.08*	**2.62**	**1.89**	0	**2.38**	*0.58*	0	*1.02*	*0.22*
G	*0.70*	**1.64**	*0.46*	*1.07*	*1.05*	**2.28**	**1.36**	0	0	*0.05*	0	0	**1.65**	0	**1.70**	**3.10**	**1.30**	*0.38*
C	*0.78*	**1.90**	0	0	*0.83*	0	*-0.38*	*0.01*	*0.09*	0	**2.00**	**2.24**	**1.70**	**2.66**	**1.50**	**2.76**	**1.52**	*0.49*

Binding of the operator O_R1 with the λ-repressor [6]

s_i	T_{-8}	A_{-7}	C_{-6}	C_{-5}	T_{-4}	C_{-3}	T_{-2}	G_{-1}	G_0	C_{+1}	G_{+2}	G_{+3}	T_{+4}	G_{+5}	A_{+6}	T_{+7}	A_{+8}
A	*0.4*	0	*0.6*	*1.0*	*-0.2*	**3.6**	**2.8**	*0.2*	**1.1**	**2.5**	**2.9**	**3.1**	**2.0**	*0.3*	0	**2.8**	0
T	0	**1.4**	*0.8*	*1.0*	0	*0.5*	0	*0.5*	**2.3**	**2.7**	**3.2**	**3.7**	0	*1.0*	*-0.1*	0	*0.4*
G	*0.6*	**1.1**	*0.7*	*0.6*	*0.4*	**3.4**	*1.0*	0	0	**2.5**	0	0	**2.9**	0	*-0.2*	**1.1**	*0.4*
C	*0.4*	*1.0*	0	0	*0.9*	0	*0.3*	*0.4*	**1.4**	0	**3.4**	**3.7**	**2.5**	*0.6*	*0.2*	**2.0**	*0.5*

INR-induction of transcription *in vivo* [7]

s_i	G_{-3}	T_{-2}	T_{-1}	A_{+1}	T_{+2}	T_{+3}	T_{+4}
A	*0.06*	*-0.22*	**2.51**	0	**0.19**	**0.19**	**0.76**
T	*0.12*	0	0	**2.51**	0	0	0
G	0	*-0.43*	**0.57**	**1.38**	*-0.06*	**0.66**	*-0.35*
C	*0.00*	*-0.05*	*-0.38*	**0.19**	**0.28**	**0.66**	*-0.32*

INR- induction of transcription *in vitro* [7]

s_i	G_{-3}	T_{-2}	T_{-1}	A_{+1}	T_{+2}	T_{+3}	T_{+4}
A	*-0.29*	**0.50**	**2.51**	0	*0.19*	**0.50**	**0.50**
T	*0.12*	0	0	**2.51**	0	0	0
G	0	*0.06*	**0.50**	**1.54**	*-0.77*	**0.88**	*0.06*
C	*-0.23*	*0.19*	*-1.06*	**1.26**	*-0.10*	**1.04**	*0.19*

c-Myb/DNA(natural, $R_1R_2R_3$) [8]

s_i	A_1	A_2	C_3	T_4	G_5	A_6	C_7
A	0	0	**3.51**	*0.57*	**4.20**	0	*0.22*
T	**3.97**	*1.31*	**3.75**	0	**4.32**	*0.05*	*0.54*
G	**4.12**	*1.54*	**3.60**	*-0.04*	0	*0.09*	**4.20**
C	**3.88**	*1.02*	0	*0.27*	**4.29**	*-0.29*	0

c-Myb/DNA (mutant $R_2(V_{103}L)R_3$) [9]

s_i	A_1	A_2	C_3	T_4	G_5	A_6	C_7
A	0	0	**2.20**	*0.95*	**1.70**	0	*0.45*
T	**1.80**	*1.00*	**2.10**	0	**1.50**	*-0.65*	*0.35*
G	**2.10**	**1.70**	**2.3**	*0.40*	0	*0.25*	**1.50**
C	**1.80**	*1.40*	0	*0.05*	**2.40**	*-0.75*	0

GCC-box/EREBP2 (natural) [10]

s_i	A_{-4}	G_{-3}	C_{-2}	C_{-1}	G_{+1}	$C+_2$	C_{+3}
A	0	**4.78**	*3.33*	*2.03*	**3.97**	*3.36*	**6.50**
T	*0.95*	**5.75**	*2.04*	**6.50**	**6.50**	*2.15*	**6.50**
G	*-0.30*	0	**6.50**	*2.50*	0	*3.30*	**6.50**
C	*-0.52*	**5.07**	0	0	**6.50**	0	0

GCC-box/AtERF-1 (remote homolog) [10]

s_i	A_{-4}	G_{-3}	C_{-2}	C_{-1}	G_{+1}	$C+_2$	C_{+3}
A	0	**2.17**	*1.45*	*1.84*	*1.73*	*1.77*	**3.79**
T	*1.35*	**3.32**	*1.77*	**1.99**	**2.93**	*0.95*	**2.93**
G	*0.63*	0	**2.05**	*1.45*	0	*1.41*	**3.12**
C	*-0.05*	**2.80**	0	0	**3.62**	0	0

Notes: i, position; s_i nucleotide (ΔΔG-matrix format [17]; horizontal is a natural variant, vertical is point mutation); **in bold**, 50% of **higher** ΔΔG (may change **specificity** of DNA/protein-recognition), *italic,* 50% of *lesser* ΔΔG (may change DNA/protein-*affinity*).

Table 2. Detected differences between point mutations influencing affinity (low ΔΔG values) and specificity (high ΔΔG), which were detected by site localization and ΔΔG-patterns in test-systems.

DNA/protein	Substitution type (A)	Site position , i (B)	N_{AB}	N_{A-B}	N_{-AB}	N_{-A-B}	α
O_R1/Cro-repressor	Substitution may disrupt	{i odd} & {| i |>1}	17	11	4	22	0.001
O_R1/λ- repressor	specificity of	{i ≥ 0}	19	9	5	19	0.0025
DNA/INR (*in vivo*)	DNA/protein recognition	{i ≥ 0}	9	2	3	7	0.025
DNA/c-Myb (natural)	(50% with higher ΔΔG)	{i- odd}	10	0	2	9	0.0005
DNA/c-Myb (mutant)		{i- odd}	10	1	2	8	0.0025
GCC-box/EREBP2		{i- odd }	10	1	2	8	0.0025
GCC- box /AtERF-1		{i- odd }	9	1	3	8	0.01

Linear correlation between ΔΔG-patterns of test-systems			Affinity		Specificity	
DNA site	Test-system #1	Test-system #2	r	α	r	α
Inr-element	Transcription *in vivo*	Transcription *in vitro*	*0.04*	*>0.25*	**0.90**	**>0.001**
c-Myb	c-Myb, natural $R_1R_2R_3$	c-Myb, mutant	**0.90**	**<0.001**	*-0.51*	*>0.10*
GCC-box	EREBP2 (natural)	AtERF-1 (homolog)	**0.81**	**<0.01**	*0.51*	*>0.10*

Table 3. Conformational and physico-chemical DNA properties used in the present study.

Property	X	Unit	Range	[n]	Property	X	Unit	Range	[n]
Bend	β	(°)	2.16 ÷ 6.74	11	Direction	D	(°)	-154 ÷ 180	14
Propeller	ω	(°)	-17.3 ÷ -6.7	12	Probability to be contacting with a protein through minor groove	P	%	1 ÷ 18	15
Twist	Ω	(°)	29.3 ÷ 39.5	13					
Tilt	τ	(°)	-0.7 ÷ 2.8	14	Mobility to bend towards major groove	μ	Rel.un.	1.02 ÷ 1.27	16

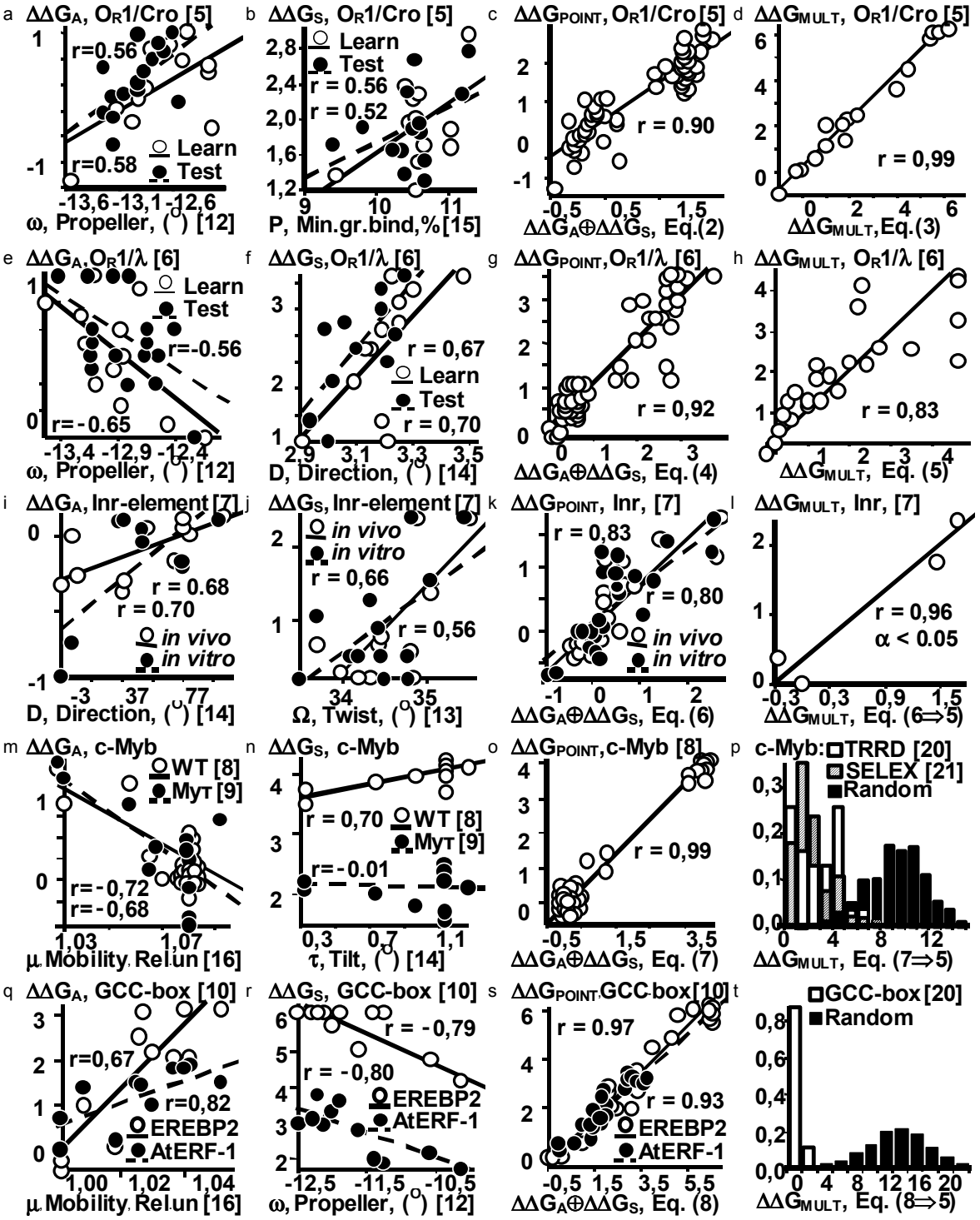


Figure 1. Experimental data on the influence of point mutations on the energy ΔΔG of DNA binding with the Cro-repressor (a-d), λ-repressor (e-h), INR-proteins (i-l), c-Myb transcription factor (m-p), EREBP-2 and AtERF-1 transcription factors (q-t).

ANALYSIS OF CONTEXT DEPENDENCIES WITHIN REGULATORY GENE REGIONS IN EUKARYOTES

Orlov Yu.L., Kosarev P.S., [1]Orlova N.G., [1]Potapov V.N.

Institute of Cytology and Genetics of SB RAS, Novosibirsk, Russia
[1]S.L. Sobolev Institute of Mathematics of SB RAS, Russia
e-mail: orlov@bionet.nsc.ru
*Corresponding author

Keywords: stochastic complexity, Markov models, genetical texts, functional sites

Resume

Motivation:

Understanding of dependencies in the context organization of regulatory gene regions in eukaryotes is of notable significance for functional marking up of *de-novo* sequenced nucleotide sequences.

Results:

By using the method of automated construction of generating contextual source-tree (a variant of a hidden Markov model) for analysis of nucleotide sequence samples containing transcription factor binding sites, we have revealed the strong dependency upon the context. Regulatory DNA sequences stored in the "Samples" database (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/) were considered by the method described.

Introduction

Promoter gene regions are the objects organized in a complex mode accordingly their context. They include different functional signals, e.g., regulatory proteins binding sites, and, in general, they are the result of super-position of several degenerate genetical codes, for example, the nucleosome code [Trifonov E., 1997]. Various regions of extensive regulatory sequence should possess by different physical-chemical properties that could be revealed statistically. With this respect, the task appears to mark off the regulatory sequence. In this procedure, both the methods aimed to detect the signals (consensusi) and to evaluate how local surroundings of these signals correspond to some predetermined statistical model are suitable. Among the examples of such models, are Markov models that are intensively applied [Peshkin L. & Gelfand M., 1999; Yada T., 1999]. So, the goals of the present study are: (i) to recognize and to mark off functional regions and to (ii) ascertain general mathematical model for their description.

Methods and algorithms

In order to study dependencies of the symbols upon the preceding context in DNA sequences, we have used the method, which was previously developed for the coding theory and data compression [Barron A. *et al*, 1997; Rissanen J., 1983]. According to our statistical model, the probability of the next in turn symbol occurrence in communication is determined by preceding context [Orlov Yu., Potapov V., 2000]. Nucleotide sequence is considered as a hidden Markov chain. The set of probabilities to generate the next symbol is determined by the state of a chain, which is unambiguously corresponds to preceding context. Such contexts may vary in length, and neither of them could be the ending of another. The maximal length of the context should be limited. Thus, for the sequences containing functional sites, the maximal context was taken as equaling to 5.

This determining set of contexts may be represented graphically in a form of a tree, which has 4 branches at each level (see Figures 1, 2). The branches correspond to nucleotide alphabet (A,T,G,C). By this tree, the set of contexts may be reconstructed, which is significant for the sequence under study, by following the route from leaves (suspended vertexes) to the root of a tree. Such tree is called a generating source-tree and together with the sets of probabilities corresponding to the contexts represents a statistical model of a sequence studied.

Markov models of the higher order may describe the sample more exactly. However, increase in the order of a model leads to appearance of numerous excessive parameters. In our approach, excessive parameters (contexts) are automatically removed during evaluation of stochastic complexity by introducing limitations on complexity (dimension) of a model [Orlov Yu., Potapov V., 2000]. Hence, the rest contexts are statistically significant for the sequence analyzed. This statistical verification makes the grounds for individual analysis of the contexts with maximal length and for construction of heuristic procedures for the functional site recognition.

Implementation and results

By the method suggested, both single sequences and groups of sequences may be analyzed, under supposition that they are generated by a single source generating the symbols. It is interesting to analyze the sequences without clearly expressed homology, but performing one and the same function. In this case, they should be similar by oligonucleotide content. We have found that non-random nucleotide contexts, which are present in all the sequences of the sample, may play a structural role for DNA sequences of the type under analysis.

We have analyzed the following samples of nucleotide sequences: 1) transcription factor binding sites stored in the «Samples» database (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/) (45 factors); 2) other types of functional sites from this database; 3) eukaryotic promoters of tissue-specific genes and "housekeeping" genes.

An analysis has revealed that transcription factor binding sites in most of the samples studied are characterized by differing structure of the source-trees (Fig.1). They vary in order of Markov chain, in tree topology, and in number of branches in the source-tree. It is seen that for one and the same site, the contexts of various lengths can be important. These data give evidence about different organization of contexts of transcription factor binding sites relatively the source generating these sequences.

In order to testify the method, we have generated the random samples of nucleotide sequences with the same length and number as the sequences in the studied samples, which were extracted from the «Samples» database. We have considered two variants of generating random sequences: 1) with fixed frequencies of single nucleotides; 2) with fixed frequencies of dinucleotides. The trees generated by the samples of such random sequences have no branches in the first case considered (dependency upon preceding context is absent), whereas in the second case, the trees have the branches only at a single level (dependency only from a single preceding nucleotide).

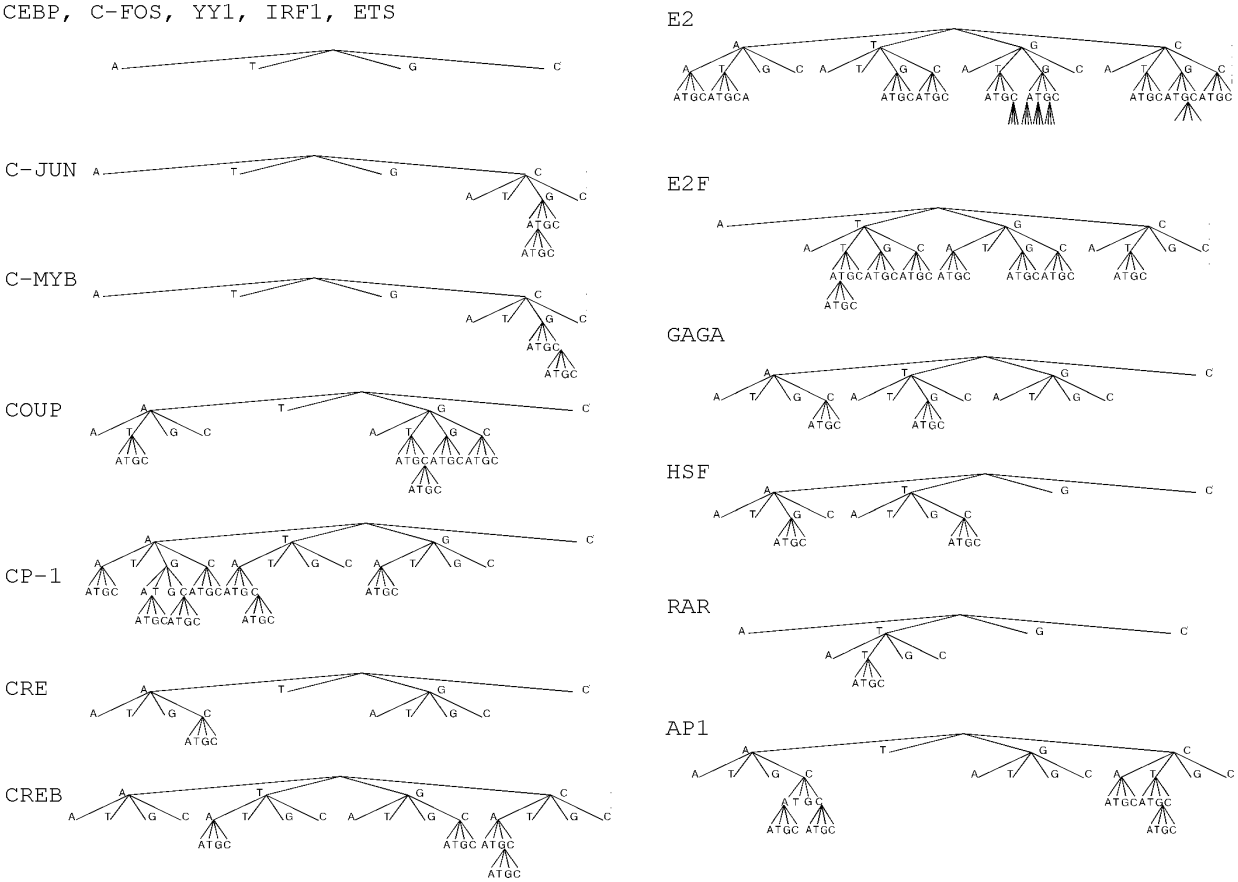


Figure 1. An example of generating source-trees for transcription factors binding sites stored in the «Samples» database (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/). The names of factors are given to the left of each tree. The sites binding transcription factors CEBP, C-FOS, YY1, IRF1, and ETS have similar structure of generating source-tree.

Hence, "unsymmetrical" branches in the source-trees constructed really correspond to non-random contexts and, therefore, they may be interpreted as the signals specific for the given types of sequences.

In general, analysis of promoter regions has revealed rather strong dependency upon the preceding context, that is, at least two-three nucleotides are significant (Markov chain of the second-third order). The reasons, why these dependencies are weaker in promoters of "housekeeping" genes (Fig. 2) are still unknown. As can be seen from the Figure, the source-tree of tissue-specific gene promoters is larger and it almost totally

incorporates the source-tree of "housekeeping" gene promoters. We can suppose that difference may be explained by chromatin structure. "Housekeeping" gene promoters as well as regions of active gene transcription should be mainly free of nucleosome packing. Figure 3 presents the context tree for the sequences by 400 bp length, containing nucleosome binding site [Ioshikhes I. & Trifonov E., 1993]. The tree for tissue specific gene promoters (Fig.2, bottom) is more similar to the "nucleosomal" tree (Fig.3).

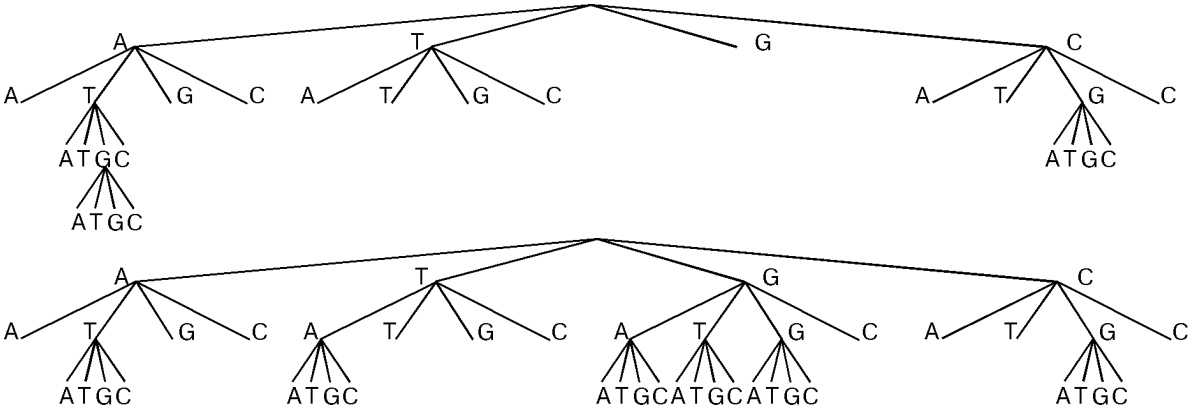


Figure 2. Generating source-trees for promoters of "housekeeping" genes (top) and tissue-specific genes (bottom). The similarity and discrepancies of trees could be noted in dependance of oligonucleotides, which contain G nucleotide for the sets of promoters classified by tissue-specificity.

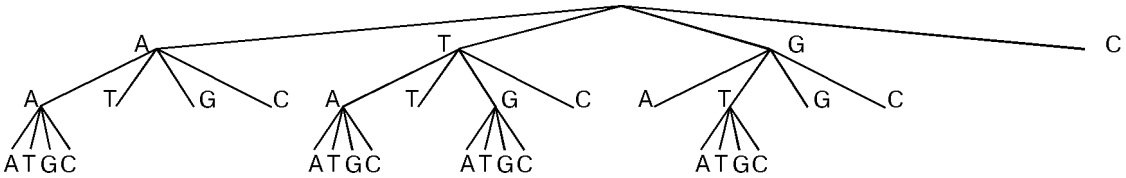


Figure 3. Generating source-trees for sequences, containing nucleosome binding site.

Discussion

The visualization of unique structure of context trees, analogous to those illustrated in Fig. 1,2, is typical for the samples considered. Empirical approach of such detection of non-random contexts supplements the estimates of under- and over-representability of oligonucleotide frequencies [Karlin S., 1995], which are characteristic for functional genome regions and genomes of various organisms. Besides, all non-random contexts can be detected, not by calculating relative occurrence of the nucleotides, but by means of ordering local organization of symbols after these contexts (relative distribution of symbols' occurrence frequencies).

Determining of dependency upon the context by the method suggested provides a basis for construction of methods necessary for recognition of particular functional regions of regulatory sequence. It was established that Markov chain (dependency upon the preceding context) for functional sites has, as a rule, the first-second order.

Acknowledgements

The authors are grateful to G. orlov for help in translation of the manuscript into English, to N. Kolchanov, E. Ignatieva, O. Podkolodnaya, V. Levitsky, D. Vorobiev, and V. Babenko for valuable comments and scientific discussion. The work was supported by Russian Foundation for Basic Research and Integration project of SB RAS.

References

1. Barron A., Rissanen J and Yu B. (1997) The minimum description length principle in coding and modelling. *IEEE Trans. Inform. Theory*, **43**, N.5, 669-683.
2. Ioshikhes I., Trifonov E.N. (1993) Nucleosomal DNA sequence database. *Nucl. Acids. Res.*, **21**, 4857-4859.
3. Karlin S. and Burge C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends in genetics*, **25**, N.7, 283-290.
4. Orlov Yu.L., Potapov V.N. (2000) Estimation of stochastic complexity of genetical texts. *Computational technologies (Novosibirsk)*, **5**, spec.issue, 5-15.
5. Peshkin L. and Gelfand M.S. (1999) Segmentation of yeast DNA using hidden Markov models. *Bioinformatics,* **15**, N.5, 980-986.
6. Rissanen J. (1983) A universal data compression system. *IEEE Trans.Inform.Theory*, **IT-29**, N.5, 656-664.
7. Trifonov E.N. (1997) Genetic sequences are the result of superposition of several codes. *Molerularnaya biologia (Moscow),* **31**, n.4, 759-767 (in Russian).
8. Yada T., Nakao M., Totoki Y., Nakai K. (1999) Modelling and predicting transcriptional units of *Escherichia coli* genes using hidden Markov models, *Bioinformatics*, 15, N.12, 987-993.

DETECTION OF CIS-ACTING REGULATORY ELEMENTS IN PLANTS: A GIBBS SAMPLING APPROACH

*[1]Thijs G., [2]Rombauts S., [1]Lescot M., [1]Marchal K., [1]De Moor B., [1]Moreau Y., [2]Rouzé P.

[1] KULeuven, Heverlee, Belgium
[2] University of Gent, Gent, Belgium
e-mail: gert.thijs@esat.kuleuven.ac.be
*Corresponding author

Keywords: motif search, statistical modelling, gibbs sampling, cis-acting element, plant promoter, microarray

Resume

Motivation:

Microarray analysis allows identifying clusters of co-regulated genes. Appropriate data analysis (pattern recognition) of promoter regions of co-regulated genes might allow gathering insight in promoter structure on a large scale. Our research aimed at developing an appropriate algorithm (based on Expectation Maximization [1] (EM) and Gibbs Sampling [2]) to extract motifs from sets of co-regulated genes. Concomitantly, a database (PlantCARE [3]) of plant cis-acting regulatory elements is constructed to store the information retrieved from our in silico motif predictions and to be able to design appropriate data sets in order to validate our algorithms.

Results:

At present, PlantCARE contains 380 distinct sites: 128 from monocots, 247 from dicots and 5 from other higher plants, describing more than 125 individual promoters. The currently available database allows:

- Retrieval of plant species-specific sites, cross-linked with additional information.
- Search for known sites in user-submitted sequences.

Our implementation of the Gibbs sampler adds the following extensions to the original algorithms of Lawrence [2]:

- An advanced background model improves the recognition of motifs in noisy sequences.
- A weighted zero or one occurrence of the motif in each sequence

Availibility:

PlantCARE is a relational database available via the URL: http://sphinx.rug.ac.be:8080/PlantCARE/. The motif sampler is still under development.

Introduction

The concept of earlier databases (e.g. TRANSFAC [5]), consists of storing any known promoter element irrespective of the source of origin. Presumably this holistic approach has resulted in an under representation of established plant regulatory elements, complicating efficient application of such databases to plant promoter analysis. Therefore we aim at constructing a local relational database which rigorously describes plant promoter cis-acting regulatory elements. To serve our purpose of motif finding we decided to adapt algorithms based on expectation Maximization [1] (EM) and Gibbs Sampling [2] since to our knowledge these algorithms constitute one of the most powerful unsupervised motif detection methods.

Methods and algorithms

PlantCARE. Every promoter element for which evidence in literature was available was annotated. For each element the corresponding degree of confidence (possible, putative, experimental) together with the nature of the evidence (experimental procedures versus homology searches) was added.

Gibbs Motif Sampling. To detect regulatory elements in a set of co-regulated genes a Gibbs sampling strategy [4] as described by Lawrence [2] was used. The resulting motif is presented as a probability matrix (motif model) $M = [q_j^l]$, j=1,..,W and l=A,C,G,T, with q_j^l the probability of finding base l at position j in the motif. The first extension to the existing algorithm is the description of an advanced background model which is based on a m^{th} order Markov chain, represented by a transition matrix $T = [t_j^l]$. In the original algorithm, the background model is described by the single nucleotide frequency q_0^l.A second extension is the assignment of a weight γ_i

to each sequence S_i. The parameter γ distinguishes sequences with a motif ($\gamma \rightarrow 1$) from the ones without motif ($\gamma \rightarrow 0$).

Algorithm

Create an initial alignment vector A={a_1,...,a_n}, describing the putative start position of the motif over the different sequences.

Initialize the probabilities of observing a motif in each sequence to an initial guess (e.g., γ=0.5).

Sample Q_i, i=1,...,N, according to the binomial distribution with parameter γ_i.

Select one sequence S_z. Use all sequence except S_z and the ones with Q_i equal to zero to calculate the motif model based on the current set of positions (A).

Assign to each segment x in S_z a weight $W_x = P_z^x / P_z^0$. P_z^x is the probability of S_z being generated given the position of segment x, the motif model and the background model. P_z^0 is the probability of S_z being generated by the background model.

Sample a new alignment vector (A) with respect to the weights W_x.

Update the weighting factor γ_z.

Repeat from 2 until a stable motif is found.

The motif is called to be stable if the probability matrix does not change significantly anymore between two iterations. In this algorithm $P_z^x = P(S_z \mid x, M, T)$ can be written as the product of the three segments: background, motif, and background.

$$P_z^x = P(S_z \mid x, M, T) = P_{bg1} P_{Motif} P_{bg2}, \text{ with } P_{Motif} = \prod_{j=1}^{W} q_j^{b_{x+j-1}}$$

$$P_{bg1} = P(b_1, ..., b_m) \prod_{j=m+1}^{x} P(b_j \mid b_{j-1} ... b_{j-m}) \text{ and } P_{bg2} = \prod_{j=x+w+1}^{L} P(b_j \mid b_{j-1} ... b_{j-m})$$

If there is no motif in S_z, the sequence is entirely generated by the background model T.

$$P_z^0 = P(S_z \mid Q_z = 0, M, T) = P(b_1, ..., b_m) \prod_{j=m+1}^{L} P(b_j \mid b_{j-1}, ... b_{j-m})$$

In step 6, we update the weight vector γ, describing for each sequence the probability to contain a motif:

$$\gamma_i = P(Q_i = 1 \mid S_i, M, T) = P(S_i \mid Q_i = 1, M, T) P(Q_i = 1 \mid M, T) / P(S_i \mid M, T)$$

If we assume that the prior P(Q_i=1|M,T) is independent of the model and the background, we can estimate P(Q_i=1) by taking the average of the current vector γ.

Data sets. To test the performance of the algorithm following data sets were used.

3. *G-box sequences*: This data set was constructed from genes known to contain a G-box in a well defined position (PlantCARE). 300 bp of the region upstream non-coding region was selected. The G-box consensus CACGTG is a rather well conserved motif.
4. *Light-induced genes*: Set of 22 co-expressed genes from one of the first Arabidopsis thaliana microarray experiments.
5. *Intergenic regions*: This data set contains only regions between two experimentally verified genes of A. thaliana. This set was used to construct a reliable transition matrix (background model).

Implementation and results

PlantCARE. The central node of the database is the 'name of site'-table (Fig. 1) , having the most general characteristics on the sites. Furthermore, a table for references is included, as literature is, for now, the most important source of information, and as the sites are organised taking the plant species into account a table for that purpose is provided.

To the central table are linked, the tables 'consensus', 'matrix' and 'site present in gene'. The latter being the table for storing specific sites on specific promoter sequences. The promoter sequences are not stored in the database as such but as an accession number in the 'genes'-table for linking with EMBL, assuring an up to date sequence at all time.

The information on factors, recognising a site can also be stored within the database. Our database has also all the necessary tables to link it to different existing databases, such as TRANSFAC [5].

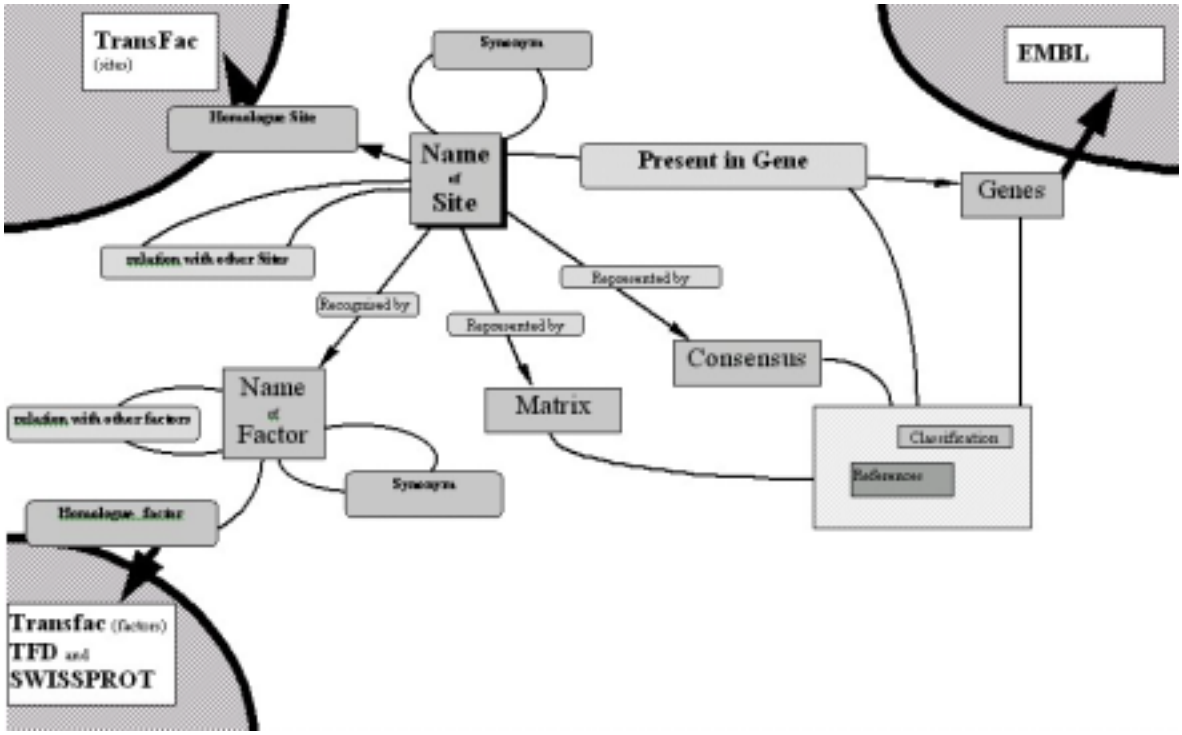


Figure 1. Scheme of PlantCARE

Motif Sampling. A first test run of the algorithm was performed on the well defined G-box data set. The initial parameters of the algorithm are weighting vector γ the motif length, the alignment vector and the background model. Most of the motifs detected by our algorithm corresponded to the G-box consensus sequences as annotated in the database. Only in 6 of the 34 sequences the wrong motif was found. For those motifs either γ_i was small or the motif probability itself was small. Moreover, the G-boxes of the corresponding sequences in PlantCARE, were annotated as putative (detected by homology).

By using distinct test runs, the sensitivity of the algorithm towards different initial conditions and the size of the dataset was analyzed. The influence of the initial alignment vector and the weighting vector on the final performance were found to be minor. The most important influence however seems to originate from the choice of an appropriate background model. A background model compiled from a large data set, independent from the set used for motif detection leads to the best results. To this end a background model was derived from the intergenic A. thaliana sequences. If there is no such dataset available, a background model can be constructed from the input data set. When a good background model is provided, a well-conserved motif can be detected in a data set of limited size (e.g. with no more than 2 sequences).

In the next step the 22 sequences from the light induced set were added to the G-box set. The algorithm is still able to find the G-box motif. In this mixed data set, the ratio of sequences with a high γ is higher in the G-box-containing sequences than in the sequences of the light induced genes. The sequences with a high weight γ are the ones containing a good match with the consensus CACGTG.

Discussion

The database PlantCARE aims at giving a rigorous description of all known cis-acting promoter elements in plants. It is being constructed as such to avoid redundancy. A relational database which takes full advantage of referential links, enables data to be retrieved in many ways allowing for instance to query the database on all the sites present on a promoter, on all the sites occurring in a particular plant species or group, or even on authors who have published on genes having particular functions associated with a specific site occurring in one plant species.

Our implementation of the motif detection algorithm allows detection of the G-box motif in noisy plant datasets. Its enhanced background model constitutes one of the major improvements. So far the algorithm allows detecting one or no occurrence of the motif per sequence. Extensions that would allow retrieving more occurrences of one motif per sequence or multiple motifs are currently under development.

Acknowledgements

Gert Thijs is a research assistant with the IWT; Yves Moreau is a post-doctoral researcher of the FWO; Prof. Bart De Moor is research associate at the FWO and professor extraordinary at the KULeuven. This work is supported by several institutions: 1. The Flemish Government: a. Research Council K.U.Leuven: Concerted

Research Action Mefisto-666. b. The FWO projects G.0240.99, G.0256.97, and Research Communities: ICCoS and ANMMM. c. IWT projects: STWW. 2. the Belgian State, Prime Minister's Office - OSTC – a. IUAP P4-02 (1997-2001). 3. Industrial Contract Research: Data4S. The scientific responsibility is assumed by its authors. We would like to thank Patrice Déhais for his help and expertise on database design. Pierre Rouzé is Research Director of INRA (Institut National de la Recherche Agronomique, France).

References

1. Bailey T. L. and Elkan C. (1995). "The value of prior knowledge in discovering motifs with MEME." *Proceedings ISMB* 3: 21-9.
2. Lawrence C. E., Altschul S.F., Boguski, M.S., Liu, J.S., Neuwald A.F., Wootton J.C. (1993). "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment." *Science* 262 (5131): 208-14.
3. Rombauts S., Dehais P., Van Montagu M., Rouze P. (1999). "PlantCARE, a plant cis-acting regulatory element database" *Nucleic Acids Res*. 27 (1): 295-6.
4. Thijs G., Moreau Y., Rombauts S., De Moor B., Rouze, P. (1999). "Recognition of gene regulatory sequences by bagging of neural networks." *Proceedings ICANN'99*, 988-993.
5. Heinemeyer T., Chen X, Karas H, Kel A.E, Kel O.V., Liebich I., Meinhardt T., Reuter I., Schacherer F. and Wingender E (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res* 27 (1): 318-22

DISCOVERY AND MODELING OF TRANSCRIPTIONAL REGULATORY REGIONS

[1]Fickett J.W. and [2]Wasserman W.W.

[1] Bioinformatics Group, SmithKline Beecham Pharmaceuticals, King of Prussia, USA
[2] Center for Genomics Research, Karolinska Institute, Stockholm, Sweden
e-mail: James_W_Fickett@sbphrd.com
*Corresponding author

Keywords: transcriptional regulatory region, regulatory network, regulatory language

Resume

A complex network of regulatory controls governs the patterns of gene expression. Enabled by the tools of molecular cloning, initial experimental queries into the gene regulatory network elucidated a wide array of transcription factors and their cognate binding sites from hundreds of genes. The recent fusion of genome-scale experimental tools, a more comprehensive gene catalog, and concomitant advances in computational methodology, has extended the range of questions being posed. The potential to further our understanding of the biochemical mechanisms of transcriptional regulation and to accelerate the delineation of regulatory control regions in the human genome is enormous.

COMPOSITE MODULES - THE DNA BLUEPRINTS OF COMBINATORIAL TRANSCRIPTIONAL REGULATION IN MULTICELLULAR ORGANISMS

[1]*Kel A.E., [1]Kel-Margoulis O.V., [1]Romaschenko A.G., [2]Wingender E., and [1]Ratner V.A.

[1]Institute of Cytology & Genetics SB RAS, Novosibirsk, Russia
e-mail: kel@bionet.nsc.ru
[2]Department of Gene Analysis, Gesellschaft für Biotechnologische Forschung mbH, Germany
*Corresponding author

Keywords: transcriptional regulation, combinatorial regulation, gene expression, composite regulatory elements, transcription factors, DNA functional sites

Resume

Motivation:

Composite modules (CMs) in transcriptional regulatory regions of eukaryotic genes are autonomous regulatory units containing two or more closely situated binding sites for distinct transcription factors. Similar CMs providing specific regulatory function could be found in regulatory regions of many coexpressing genes. Such CMs could be used as explicit benchmarks for the regulatory regions of tissue or developmental stage specific genes expressed in certain cellular conditions. Therefore search of potential CMs is desired for functional annotation of long purely characterised genomic sequences.

Results:

Promoters of different gene sets coexpressed during realization of specific programs of cell ontogenesis were analyzed: (i) genes expressed in the activated T-, B- lymphocytes, that is upon cell cycle entry from the quiescent state; (ii) genes with maximal level of expression at the G1/S transition and in the S-phase of cell cycle in the various types of cells; (iii) genes specific for terminally differentiated cells (brain – specific genes as examples). We have found that promoters of these gene sets are imprinted by distinct composite modules that are responsible for the principal component of expression pattern of these genes. We have revealed complex structure of immune cell-specific CMs that may encode multiple alternative arrangement of transcriptional regulatory complexes in different cellular situations.

Availability:

The software has been developed for identification of functionally related sets of promoters in the course of genome functional annotation (for NFAT and E2F CMs as examples) (http://compel.bionet.nsc.ru/FunSite.html).

Introduction

The fundamental block-hierarchical principle of organization of molecular genetic systems (Ratner, 1990) realized in apparent modular structure of genomic DNA, both in its structural and regulatory parts (Ratner, 1992). In recent years much attention has been paid to the investigation of the modular structure of regulatory regions that control transcription of eukaryotic genes (Dynan, 1989; Johnson & McKnight 1989; Werner, 1999; Struhl, 1999). This is a very important principle for understanding molecular mechanisms of functioning of these regions, their evolution and what is particularly important for deciphering complex mechanisms of differential gene expression in multicellular organisms.

Realisation of diverse intracellular programs in different tissues and cell types, stages of development, phases of cell cycle and other cellular situations is carried out by the precisely organised differential expression of specific sets of genes.

Such regulation is provided by precisely organised binding of a multiplicity of transcription factors (TFs) to their target sites (cis-elements) in regulatory regions. Cis-element combinations provide a structural basis for the generation of unique patterns of gene expression. Co-ordinated expression of specialised sets of genes is achieved by the presence of similar combinations of cis-elements that could recruit specific sets of transcription factors. These TFs interact with each other and with particular components of the basal transcription complex as well as with coactivators/corepressors, histone acetylases/deacetylases, therefore making up function-specific multiprotein complexes.

Combinations of two or more TF binding sites that could be found in regulatory regions of similarly regulated genes will be referred to as composite modules (CM). The minimal CMs are the composite elements consisting of two TF binding sites. We collect information on structure and function of known composite elements and systematize them in the COMPEL database (Kel et al., 1995; Kel-Margoulis et al., 2000).

The hierarchy of universal functional blocks in transcription regulatory regions of eukaryotic genes becomes evident now (Heinemeyer et al., 1998; Klingenhoff et al., 1999; Werner, 1999). According to the hierarchy, TF binding sites correspond to the minimal functional blocks of transcription regulatory regions. CMs of different complexity correspond to the next levels of functional block hierarchy. They provide a basis for combinatorial regulation of transcription, for coordinated and alternative gene regulation.

In the present paper we study various composite modules that may contain two or more TF binding sites as well as additional contextual features specific for groups of coordinately regulated genes.

Results and Discussion

Composite modules specific for gene functional groups

We have analyzed promoters of the three groups of genes whose expression is maximal in the course of realization of different cell-cycle programs. Gene sets under study were: (i) genes expressed in the activated T-, B- lymphocytes that correspond to the cell cycle entry from the quiescent state (***T-promoters***); (ii) genes with maximal level of expression at the G1/S transition and in the S-phase of cell cycle in various cell types (***C-promoters***); (iii) genes specific for terminal differentiation programs (e.g. ***brain***-specific genes).

We have performed a search for a number of transcription factor (TF) binding sites known to be important for regulation of gene sets mentioned above. These binding sites include E2F, NF-Y, c-Myc, Sp1, NF-AT, AP-1, NF-κB, GATA, CRE-BP1, Oct. New search methods based on dinucleotide weight matrices and facultative motif distributions were applied for high quality recognition of these sites. For revealing the composite modules (CMs) we applied several AI approaches, such as: genetic algorithms and fuzzy calculations.

We have found that the promoters under consideration are imprinted by specific composite modules containing several potential TF binding sites as well as other contextual motifs. ***T-promoters*** appeared to be highly enriched by clusters of composite elements (CEs) NF-AT/AP-1 (Kel et al., 1999). These CEs are located in close proximity to the transcription start site and often found to be the most conserved part of the promoters (Kel et al., 1999). New modules of the composite structure were revealed in ***C-promoters***. E2F sites in arrangement with Sp1 sites and additional motifs such as *GSG* and *WWTT* are highly specific for ***C-promoters*** (10 times more frequent than in EPD promoters and 100 times than in exons) (Kel and Kel-Margoulis, 1999). The difference between ***C-promoters*** and EPD promoters by the frequency of potential CMs of this type is most significant within the region [-50 to +1]. The characteristic features of ***brain***-specific promoters are extended CMs containing 3-4 sites for the ubiquitous inducible transcription factor CREB and some binding sites for developmental-stage specific TFs. This analysis has demonstrated that promoters of functionally interrelated groups of genes are characterized by high frequency of a particular CM specific for each group of genes. The results suggest that CMs are essential regulatory structures that provide the expression profile of these gene sets in the course of realization of different cell-cycle programs.

We have developed software for identification of functionally related sets of promoters in the course of genome functional annotation (for NFAT and E2F CMs as examples) (http://compel.bionet.nsc.ru/FunSite.html).

Alternative composite modules as a basis for switching programs of cell ontogenesis

Different CMs often overlap thus enabling distinct alternative regulatory complexes to assemble on the same transcription regulatory regions under different cellular conditions (Kel et al, 1995; Fry & Farnham 1999). There are some known examples when two different TFs could alternatively bind to overlapping DNA sites thus providing different features of regulation of the corresponding gene or even making a switch between alternative programs of gene expression.

Many of these examples were collected in the COMPEL database as antagonistic composite elements. Interesting examples are known for binding sites of NFAT transcription factors. For instance, it is known that factors other than NFATp/c family members can bind NFAT motifs. Transcription factors Elf-1, a member of the Ets family, and NFATp were shown to activate mouse the GM-CSF promoter through the same site on DNA and, moreover, both factors act cooperatively with AP-1 (Masuda et al., 1993; Wang et al., 1994; COMPEL, C00108 and C00081). The fact is now well known that NFATp (or related factors) can bind to the 3' half of the NF-κB consensus motif (Sica et al., 1997). This may lead to the mutually exclusive binding, depending on both nucleotide content and cellular situation. Alternative binding of NF-κB or NFATp to the proximal promoter of human IL-4 (Casolaro et al., 1995) adversely affect transcription of the IL-4 gene.

It is likely that several other transcription factors can bind cis-elements containing GGAAA that is the core motif for NFATp/c family factors. To demonstrate that this motif appears frequently among various binding sites we have counted all possible pentanucleotides in the sequences of binding sites for vertebrate transcription factors

collected in the TRANSFAC database. 1792 available sites were scanned and the most frequent pentanucleotides were selected. The GGAAA/TTTCC motif representing NFATp/c core sequence is among the 3 most frequent pentanucleotides: more than 8% of all TRANSFAC sites contain this element.

In addition, we have checked how often NFATp/c sites overlap with binding sites for other families of transcription factors. This analysis was based on the databases TRANSFAC and TRRD. We scanned the sequences of sites different from NFATp/c with the NFATp weight matrix using a threshold $q_{NFAT} > 0.87$ (Kel et al., 1999) and recorded all matrix matches that overlapped with the core of these binding sites. We found that many real sites that are targets for various factors are matching the NFATp/c weight matrix (see Tabl. 1).

Table 1. Scanning of binding site collections for different transcription factors with the NFATp/c weight matrix.

Transcription factor family	Number of sites in the set	Number of sites matching NFATp/c matrix ($q_{NFAT} > 0.87$)
NFATp/c	41	33 (80.5%)
IRF	13	7 (53.8%)
ETS	15	4 (26.7%)
C/EBPβ	23	4 (17.4%)
NF-κB	30	4 (13.3%)
E2F	26	3 (11.5%)
GATA	50	2 (4.0%)
NF-1	102	4 (3.9%)
SRF	23	1 (4.0%)
SP1	125	1 (0.8%)

Members of several transcription factor families could potentially bind to the same DNA sites in the gene regulatory regions. Such cis-elements – targets for concurrent TFs – may be the parts of different alternative composite modules. Alternative CMs may serve as a basis for the olecular mechanism of switching between different programs of cell ontogenesis through turning on or turning off the sets of genes specific for definite cellular situations.

"Fuzzy puzzle" – hypothesis of multipurpose structure of eukaryotic promoters

The multiplicity of the cellular conditions in which eukaryotic genes should be expressed is the cause of polyfunctionality of the structure of their transcription regulatory regions. We believe that this polyfunctionality is governed through alternative CMs.

Here we present a "fuzzy puzzle" hypothesis of coding multiple regulatory messages in the same DNA sequence. The structure of regulatory sequences on one hand and the specific features of transcription factors on the other hand provide a possibility to encode several regulatory programs within one regulatory region (see Fig. 1). It is known that each transcription factor has the ability to bind to a variety of different DNA sites. This is maintained by flexible mechanisms of DNA-protein interactions, when DNA conformation rather than the particular sequence context often the major role in selection of DNA targets. In addition, the ability of TFs to operate through a so-called "induced fit" mechanism (when a TF becomes finally structured only upon interaction with DNA; Frankel and Kim, 1991) greatly relaxes the restrictions from binding to various DNA sites. Besides that, the protein-protein interactions between different transcription factors in the multiprotein regulatory complex become very important. Protein-protein interactions could stabilize some low-energy protein-DNA contacts thus additionally widen the variety of target sites for particular transcription factors. The huge diversity of transcription factors functioning in the living cells multiplied by the wide choice of target sites for each TF give rise to a precondition to form multiple alternative DNA-protein complexes on the same gene regulatory region. As a result extremely complex patterns of gene expression are observed.

The "fuzzy puzzle" model that is based on consideration of composite modules is the logical development of the block-hierarchical principle of organization of regulatory genomic sequences. "Fuzzy puzzle" is the result of genome evolution of the multicellular organisms on the way of breaking the evolutional limitations caused by the requirement of multiple ontogenetic programs to be encoded in a single genome.

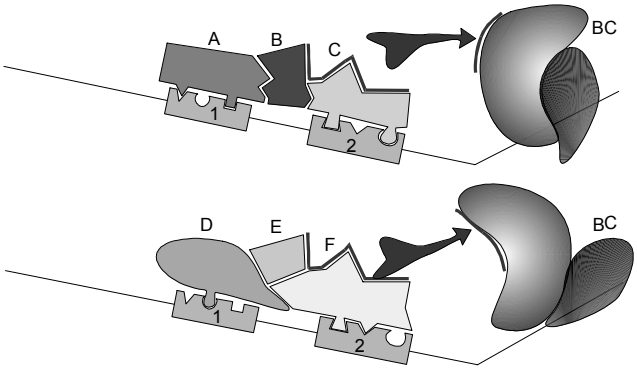


Figure 1. Fuzzy puzzle hypothesis of the multipurpose structure of the eukaryotic promoters: of coding multiple regulatory messages in the same DNA sequence. A,B,C and D,E,F – two sets of TF; 1,2 – two sites in DNA; BC – basal complex.

Acknowledgements

Different parts of this work were funded by the German Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (project no. 0311 640 and X224.6), by the

Russian Ministry of Sciences and the Siberian Branch of Russian Academy of Sciences, by the North Atlantic Treaty Organisation (grant no. 951149), and by Volkswagen-Stiftung (I/75941).

References

1. Casolaro V., Georas S.N., Song Z., Zubkoff I.D., Abdulkadir S.A., Thanos D., and Ono S.J. (1995) *Proc. Natl. Acad. Sci. U S A*, **92**, 11623-11627.
2. Dynan W.S. Modularity in promoters and enhancers. // Cell. 1989. V.58. P.1-4.
3. Gottschalk L.R., Giannola D.M. and Emerson S.G. (1993) *J. Exp. Med*; **178**, 1681-1692.
4. Frankel A.D. and Kim P.S. (1991) Modular structure of transcription factors: Implication for gene regulation. *Cell*, **65**, 717-719.
5. Fry J.Ch., Farnham P.J. (1999) Context-dependent transcription regulation. *J. Biol. Chem.*, **274**, 29583-29586.
6. Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny, N. L. & Kolchanov, N. A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.* **26**, 362-367.
7. Johnson P.F. and McKnight S.L. Eukaryotic transcriptional regulatory proteins. // Annu. Rev. Biochem. 1989. V.58. P.799-839.
8. Kel,O.V., Romaschenko,A.G., Kel,A.E., Wingender,E., and Kolchanov,N.A. (1995) *Nucleic Acids Res.*, **23**, 4097-4103.
9. Kel A., Kel-Margoulis O., Babenko V., Wingender E. Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells (1999) *J. Mol. Biol. 1999*, **288**, 353-376.
10. Kel A. and Kel-Margoulis O. (1999) Proceedings of the workshop, Bridging the gap between sequence and function. Cold Spring Harbor Laboratory, September 7-9, P.14.
11. Kel-Margoulis,O.V., Romaschenko,A.G., Kolchanov,N.A., Wingender,E., and Kel,A.E. (2000) *Nucleic Acids Res.*, **28**, 311-315.
12. Klingenhoff, A., Frech, K., Quandt, K., Werner, T. (1999). Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* **15**, 180-186.
13. Maggirwar S.B., Harhaj E.W., and Sun S.C. (1997) *Mol. Cell. Biol.*, **17**, 2605-2614.
14. Masuda E.S., Tokumitsu H., Tsuboi A., Shlomai J., Hung P., Arai K., and Arai N. (1993) *Mol. Cell. Biol.*, **13**, 7399-7407.
15. Ratner V.A. (1990) Towards the Unified Theory of Molecular Evolution (TME). *Theor. Popul. Biol.*, **38,** 233-261.
16. Ратнер В.А. (1992) *Генетика*, **28**, 5-24.
17. Ratner V.A. (1992) *Genetica (Mosk)*, **28**, 5-24.
18. Struhl K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1-4.
19. Sica A., Dorman L., Viggiano V., Cippitelli M., Ghosh P., Rice N., and Young H.A. (1997) *J. Biol. Chem.*, **272**, 30412-30420.
20. Wang C.-Y., Bassuk A.G., Boise L.H., Thompson C.B., Bravo R. and Leiden J.M. (1994) *Mol. Cell. Biol.*; **14**, 1153-1159
21. Werner, T. (1999). Models for prediction and recognition of eukaryotic promoters. *Mammalian Genome* **10**, 168-175.

FINDING TRANSCRIPTION FACTOR BINDING SITES IN COREGULATED GENES BY EXHAUSTIVE SEQUENCE SEARCH

Kielbasa Sz.M., Korbel J.O., Beule D., Schuchhardt J., and *Herzel H.

Institute for Theoretical Biology, Humboldt University, Berlin,Germany
e-mail: h.herzel@itb.biologie.hu-berlin.de
*Corresponding author

Keywords: regulatory sequences, upstream regions, exhaustive search, Ras pathway

Resume

Motivation:

Growing amounts of gene expression data provide the possibility of finding coregulated genes by clustering methods. By analysis of the promoter regions of these genes, rather weak signals of transcription factor binding sites may be detected [Zhang, 1999]. We compare existing programs and own software on yeast clusters. Therefore, we introduce the new algorithm ITB, an *integrated tool* for *box* finding, which exhaustively analyses regular expression-like patterns in promoter sequences, allowing gaps and the matching of more than one base at any position within the candidates. The applicability of ITB to predict transcription factor binding sites in human promoter sequences is evaluated.

Results:

Three publicly available algorithms were compared to our program, particularly on yeast clusters. Moreover, ITB was tested on promoter sequences of coregulated human genes. ITB is capable of predicting several verified transcription factor binding sites in yeast.

Availability:

The program ITB is available upon request.

Introduction

Basic molecular biological processes are regulated by the specific interaction of proteins and short DNA sequences. Two different approaches are used to predict transcription factor binding sites: Exhaustive analyses of oligonucleotide frequencies, for instance, as described by van Helden *et al.* [van Helden,J. et al., 1998] and non-exhaustive optimization approaches using weight matrices, like Gibbs sampling, which was first described by Lawrence et al [Lawrence,C.E. et al., 1993].

The program RSA-tools-oligo-analysis [van Helden,J. et al., 1998] compares the frequency of conserved words in a given set of promoter sequences to the frequency of those words in a training set. This method is sensitive in detecting conserved words, which are slightly over-represented in the given coregulated sequence set. Unfortunately, regulatory elements not having a conserved core sequence cannot be detected by this method. Weight matrix based methods like Gibbs sampling [Lawrence,C.E., et al., 1993] or MEME [Bailey,T.L. & Elkan,C., 1994] can predict elements without a conserved core. However, for a small number of sequences in the coregulated set, weight matrices are of limited use. Moreover, signals provided by DNA regulatory elements involved in transcription are weak and rather poly-(A), (T), or GC-rich regions of the promoter might be aligned by weight matrix-based prediction methods.

Methods and algorithms

Three publicly available tools were compared to our program ITB: The Gibbs Motif Sampler (developed by E.C. Rouchka & B. Thomson based on the work of Lawrence [Lawrence C.E., et al., 1993], RSA-tools-oligo-analysis [van Helden,J. et al., 1998], and MEME version 2.2 [Bailey,T.L. & Elkan,C. 1994]. RSA-tools-oligo-analysis cannot be run with human DNA, since the scoring method of that program is dependent on the training set and human training sets cannot be chosen in the web interface of the program. Instead, we apply ITB, a program developed by our group, which is similar to RSA-tools-oligo-analysis when run in the "ACGT"-mode. For the analysis of *Saccharomyces cerevisiae*, ITB is trained with a collection of all 5'-UTRs of yeast of length 800, while all 271 human promoter sequences obtained from EPD, the eukaryotic promoter database [Perier,R.C. et al., 1998], are used as a training set for the analysis of human DNA.

Our program written in C++ compares frequencies of conserved elements in the given promoter set to the expected frequencies of these elements, which are estimated based on the training set by using Markov chain models of varying orders. The program performs an exhaustive search – scores are calculated for all possible

6-mers built from the alphabets 'ACGT' or 'ACGTWRKSYMN' (The meaning of these symbols used in the "extended mode" of the program are: W=A or T, R=A or G, K=G or T, S=C or G, Y=C or T, M=A or C, and N=any of ACGT.), depending on the mode requested. The score expresses, in the logarithmic scale, how improbable it is for the background model (with "all motifs equally distributed") to generate the observed number of occurrences of a motif in the coregulated set (formulae from [van Helden,J. et al., 1998]), with modifications for the extended alphabet). Finally, a list of the motifs with the best scores is created, containing the most over-represented patterns in comparison to the training set. The last step, which we currently achieve "by eye", is the removal of self-overlapping patterns (like AAARAA or ATAWAT) from the list.

We compare existing methods and own software on yeast clusters and evaluate the applicability of our algorithm for human promoter sequences. Zuber J. reported the coregulation of several genes of the H-Ras signal transduction pathway [Zuber,J., et al., 2000]. They kindly provided five promoter regions of those genes. Moreover, four promoter sequences of genes upregulated by Myc [Coller,H.A. et al., 2000] were extracted from EPD [Perier R.C. et al., 1998].

Implementation and results

Coregulated gene sets of yeast used in our analysis are identical with sets used by van van Helden,J. [van Helden,J. et al., 1998]. Table 1 lists human gene sets used in the analysis. Table 2 shows the results of a preliminary analysis (without adjusting/optimizing parameters) of the performance of MEME, Gibbs sampler, RSA-tools-oligo-analysis, and ITB in upstream regions of coregulated yeast genes ("gene families"). ITB predicted most previously characterized elements correctly (top scoring element). For the GCN family, the previously characterized element was ranked at position two. MEME and the Gibbs sampler failed to predict some of the previously characterized elements. The programs were compared in a single run. Table 3 shows the type of motifs predicted by ITB run in the "extended-mode" and their rank based on the score.

Table 1.The genes used in analyses of human promoter regions.

Human protein	Genes regulated via the human protein
Myc	EP11114 EP15041 EP36018 EP37001
H-Ras	LOX LOXL1 LOXL2 RIL TSP1

Table 2. Exhaustive methods and non-exhaustive, optimizing approaches are used to predict hexanucleotides, as parts of previously characterized regulatory elements of yeast. An 'x' indicates that the previously characterized element is correctly predicted (ranked on the top position) by the program.

Gene family	MEME	Gibbs sampler	RSA-tools-oligo-analysis	ITB ("ACGT-mode")	ITB ("extended-mode")
NIT	–	x	x	x	x
MET	–	–	x	x	x
PHO	x	x	x	x	x
PDR	x	–	x	x	x
GCN	x	–	x	x	–

Table 3. Highly ranked hexanucleotides as computed by the program ITB are similar to the consensus sequences of previously characterized sites taken from TRANSFAC [Wingender *et al.*, 1996][1] or from publications by Katzmann D.J. [Katzmann,D.J. et al., 1996][2], and Kuras L. [Kuras,L. et al. 1996][3].

Family	Previously characterized element	Element predicted by ITB	Inf. cont.	Score	Rank
NIT	GATAAG[1]	KATMRS	8.1	25.6	1
MET	TCACGTG[3]	CMCRYR	9.0	29.2	1
PHO	CACGTKNG[1]	AYKWGS	8.4	25.0	1
PDR	TTCCGCGGAA[2]	CCRYGG	12.3	33.7	1
GCN	RTGACTCATNS[1]	AYGACK	11.6	15.1	2
Coregulated via					
Myc	CACGTG[1] (binding core of Myc)	MMCKKG			25
Ras	not characterized yet	SYSTST			1

Discussion

Although the first results in yeast are promising, in order to predict motifs in human gene sets we are extending the program mentioned above. Currently we develop an algorithm, which is aware of self-overlapping words and allows an exhaustive search for patterns containing gaps. Moreover, we need more human promoters of coregulated genes as well as better training sets. Since only a minority of human promoters have been experimentally verified, promoter prediction algorithms could be used (reviewed by Fickett & Hatzigeorgiou, 1997) to obtain more sequences.

References

1. Bailey,T.L. and Elkan,C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.
2. Coller,H.A., Grandori,C., Tamayo,P., Colbert,T., Lander,E.S., Eisenman,R.N., and Golub, T.R. (2000). Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. *Proc. Natl. Acad. Sci. USA*, 97, 3260-3265.
3. Fickett,J.W., and Hatzigeorgiou,A.G. (1997). Eukaryotic promoter recognition. *Genome Res.*, 9, 861-878.
4. Katzmann,D.J., Hallstrom,T.C., Mahe,Y., and Moye-Rowley,W.S. (1996). Multiple Pdr1p/Pdr3p Binding Sites Are Essential for Normal Expression of the ATP Binding Cassette Transporter Protein-encoding Gene. J. Biol. Chem., 271, 23049-23054.
5. Kuras,L., Cherest,H., Surdin-Kerjan,Y., and Thomas,D. (1996). A heteromeric complex containing the centromere binding factor 1 and two basic leucine zipper factors, Met4 and Met28, mediates the transcription activation of yeast sulfur metabolism. EMBO J., 15, 2519-2529.
6. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F., and Wooton,J.C. (1993). Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science, 262, 208-214.
7. Perier,R.C., Junier,T., and Bucher,P. (1998). The eukaryotic promoter database EPD. Nucleic Acids Res., 26, 353-357.
8. van Helden,J., Andre,B., and Collado-Vides,J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J. Mol. Biol., 281, 827-42.
9. Wingender,E., Dietze,P., Karas,H., and Knuppel,R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res., 24, 238-241.
10. Zhang,M.Q. (1999). Large-scale gene expression data analysis: a new challenge to computational biologists. Genome Res., 9, 681-688.
11. Zuber,J., Tchernitsa,O.I., Hinzmann,B., Schmitz,A.C., Grips,M., Hellriegel,M., Sers,C., Rosenthal,A., and Schafer R. (2000). A genome-wide survey of RAS transformation targets. Nat. Genet., 24,144-52.

KERNEL METHOD FOR ESTIMATION OF FUNCTIONAL SITE LOCAL CONSENSI. CLASSIFICATION OF TRANSCRIPTION INITIATION SITES IN EUKARYOTIC GENES

[1]*Tikunov Y., [2]Kel A.

[1]United Institute of Geology, Geophysics & Mineralogy SB RAS, Novosibirsk, Russia
e-mail: tikunov@uiggm.nsc.ru
[2]Institute of Cytology & Genetics SB RAN, Russia
e-mail: kel@bionet.nsc.ru
*Corresponding author

Keywords: transcriptional regulation, DNA functional sites, cluster analysis, weight matrixes, start of transcription

Resume

Motivation:

DNA sites bearing the similar functions in different genes often differ significantly by their nucleotide sequences. In many cases a set of binding sites can bedivided into subgroups that tend to be bound preferably by a particular member of a transcription factor family. Identification of such subgroups is essential for developing precise methods of binding site recognition in genomic sequences and for understanding peculiarities of transcriptional regulation mechanisms. In order to detect such subgroups in the sets of DNA functional sites it is necessary to create new clusterization methods for revealing common sub-sequences within each subgroup of sites (local consensi and corresponding weight matrixes).

Results:

We have developed a novel clusterization method for the analysis of the sets of functional sites based on kernel estimation. The method developed has been applied for the analysis of transcription initiation sites in eukaryotic genes. We have revealed two local consensi that correspond to different functional and taxonomic groups of genes.

Availability:

The program for estimation of local consensi is available upon request. Two local consensi found for transcription initiation sites could be used for scanning sequences by applying MonoMatch program available at http://compel.bionet.nsc.ru/cgi-bin/MoMatch/momatch.pl. Please, select the profile: “TranscriptionStart”.

Introduction

One of the important problems of eukaryotic genome analysis is recognition of functional sites providing control of the main molecular genetic processes in the cell, in particular transcriptional regulation. Extraordinary diversity of regulatory sites is the main complication for investigation of site structure. Frequently sites with the similar function in different genes differ significantly from each other by the nucleotide sequences. This phenomenon could be explained both by different ways of their origin and by the multiplicity of molecular mechanisms their functioning. It is known that the same site can be a target for numerous transcription factors different by the structure, functional properties as well as by their ability to co-operate with other factors. On the other side, a subgroup of binding sites specifically recognized by a particular member of a transcription factor family may be revealed within the set of binding sites of similar structure. For detection such subgroups (clusters) in the sets of DNA functional sites we have developed a novel method based on kernel estimation.

Currently, there are a number of methods for the cluster analysis: method of k-means, methods for dendrogram construction among others. These methods are widely applied in bioinformatics, in particular in the field of molecular evolution (R. Durbin et al, 1998). In the recent years the methods of cluster analysis have been successfully used for the analysis of the splice sites (Kudo et al., 1992; Rogozin and Milanesi, 1997), transcription factor binding sites (Kel et al., 1995; Kolchanov et al., 1998); for the analysis and recognition of eukaryotic promoters (Kondrakhin et al., 1995).

However, as it was admitted by Aiwasyan and his colleagues (1989), the majority of existing methods of cluster analysis are rather heuristic and results of clusterisation sometimes are difficult to interpret. In the present paper we are dealing with the statistical basis of clusterisation by applying smooth kernel estimations of weight matrixes for separate subgroups of sites in a space of contextual features. Search of local consensi is based on the revealing specific points that are characterised by the maximal frequency in the close proximity within the space of features.

We have applied the method developed for the analysis of transcription initiation sites in eukaryotic genes. We have revealed two local consensi and estimated two correspondent weight matrixes. Using these weight matrixes we have sorted all promoters (Eukaryotic Promoter Database, release 56) into the three groups and found that they coincide with different functional and taxonomic groups of genes.

Method

Let a set of letter sequences of fixed length be a result of a random experiment. Let every separate sequence be an element of space ***L***. There may be groups of sequences in space ***L*** which are situated in some compact regions of space ***L***. Each group is characterised by its own consensus and its own weight matrix $||p_{jl}||$ (index *jl* means *l* letter in *j*-th position). In order to get the weight matrix estimation for given local consensus we set a weight W_L for every sequence using the following equation:

$$W_L = (c_p \cdot P_L)^{1/h} \qquad (1)$$

here, P_L is expected probability of sequence *L* in the neighbourhood of the consensus (P_L is calculated from matrix $||p_{jl}||$); *h* is smoothing parameter defined as a constant; c_p is a normalisation factor. Thus, function W_L is a weight kernel which is the basis for estimation of every separate local weight matrix $||p_{jl}||^*$. Matrix elements are calculated as estimations of weighted probabilities of letters in the sequences in accordance with weights defined by kernel W_L. Assuming $||p_{jl}||^* = ||p_{jl}||$ we get the following system of equations:

$$R_L = \sum_j \gamma_{jl} \qquad (2.1)$$

$$s_{jl} = \sum_{L \in \boldsymbol{L}^*_{jl}} e^{-(R_L - \gamma_{jl})/h} \qquad (2.2)$$

$$s_{jl_0} = \max_j (s_{jl}) \qquad (2.3)$$

$$\gamma_{jl} = \ln\left(\frac{s_{jl_0}}{s_{jl}}\right) \qquad (2.4)$$

here, R_L is the distance from the sequence *L* to the given local consensus; γ_{jl} is distance coefficient for *l* letter in *j*-th position; s_{jl} is weighted sum for *l* letter in *j*-th position. Thus if there are several groups of sequences in space ***L*** and every group is characterised by its own local consensus and weight matrix then a separate solution of the equation system (2) for every group should exist. It should be pointed out that solutions of equation system (2) correspond to extremes of the following functional Φ_W from the function W_L

$$\Phi_W = \frac{\sum_{L \in \boldsymbol{L}^*} W_L}{\left(\sum_{L \in \boldsymbol{L}} W_L^{1+h}\right)^{1/(1+h)}} \qquad (3)$$

here, $\boldsymbol{L}^*$ is the considered set of sequences. Local maximums of functional Φ_W correspond to the regions of space ***L*** characterised by the highest density of sequences or in other words local maximums correspond to clusters of sequences.

The algorithm of searching the local consensi is based on the mathematical model described above. The value for smoothing parameter *h* was chosen such that to provide certain stability of solutions. This depends on the sequence length, size of the alphabet, number of sequences in the set and distribution of their frequencies. We use the following iterative procedure for searching local consensi:

1. Initialisation of the algorithm by setting the initial values γ_{jl}. For that we select a sequence *L* and set γ_{jl} = 0, where *l* is a letter in *j*-th position of sequence *L*. All other values of γ_{jl} set to 1.
2. Calculation of distance R_L in accordance with formula (2.1).
3. Calculation of partial sums s_{jl} in accordance with formula (2.2).
4. Determination of maximal values of s_{jl_0} for every position in accordance with formula (2.3).
5. Calculation of new values γ'_{jl} in accordance with formula (2.4).
6. Comparison of new calculated values γ'_{jl} with the previous values γ_{jl}. If the difference between coefficients is higher then a predefined value then the new calculated values γ'_{jl} are put instead of γ_{jl} and procedure is repeated from point 2. Otherwise the iterations are interrupted and we get a solution.

By starting the algorithm in turn from every sequence in the set we get a number of non-coincident solutions. Every such solution corresponds to a separate local consensus and to a sub-group of sequences (cluster). This group is characterised with its own weight matrix $||p_{jl}||$ (calculated from γ_{jl} values).

Results

Analysis of transcription initiation sites.

A set of eukaryotic transcription initiation sites comprising 836 sequences (EPD, release 56) was taken for the analysis. Nucleotide sequences of all promoters were extracted from EMBL databank and aligned in respect to the transcription start point. For further analysis we have used parts of this alignment located in a sliding window of the length from 5 up to 40 bp. All possible locations of both windows in the region –20/+20 were analysed. Varying the smoothing parameter h , we searched for the local consenci and examined the number of found solutions. Stable solutions that did not change considerably under essential variation of smoothing parameter were the most interesting for us.

We have found that solutions obtained are the most stable in the region [-17 , +2]. In this region we have obtained two stable local consensi that are characterised by weight matrixes shown in the Table 1. One can see that one of these consensi preferably contains the letters G and C and the other contains C and T. It should be pointed out that both consensi contain the most conservative CA dinucleotide just before the transcription start point.

Table 1. Local consensi and corresponding weight matrixes for the region [-17 , +2] in respect to the transcription start. Maximal values of weights for every position are boldfaced. The local consensus sequences are shown below each table.

Local consensus N1 (GC)

	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2
A	.19	.11	.19	.12	.12	.09	.13	.11	.12	.17	.09	.19	.14	.22	.07	.07	**.60**	.09	.16
G	.37	.46	.50	.53	.43	.37	.36	.45	.42	.48	.47	.36	.52	.26	.41	.15	.22	.35	.18
C	.34	.29	.20	.24	.37	.26	.41	.32	.33	.21	.33	.31	.23	.41	.29	**.67**	.12	.33	.26
T	.10	.14	.12	.11	.09	.28	.10	.12	.13	.15	.10	.13	.11	.12	.22	.11	.06	.23	.39
	g/c	G	G	G	G	G	C	G	G	G	G	g	G	C	G	C	A	g/c	t

Local consensus N2 (CT)

	-17	-16	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	+1	+2
A	.17	.13	.21	.16	.13	.11	.11	.12	.16	.15	.13	.18	.15	.15	.11	.07	**.60**	.12	.15
G	.25	.26	.23	.24	.33	.18	.21	.19	.23	.23	.22	.19	.20	.19	.18	.11	.12	.19	.10
C	.45	.47	.36	.43	.28	.33	.37	.32	.24	.41	.34	.23	.46	.41	.23	**.61**	.14	.46	.45
T	.13	.13	.20	.17	.26	.38	.31	.37	.38	.21	.31	.40	.19	.25	.48	.20	.14	.23	.30
	C	C	C	C	g/c	T	c	t	t	C	C	T	C	C	T	C	A	C	C

All promoters from EPD were classified on the basis of these two weight matrixes. Distances to both local consensi (GC and CT) were calculated for every promoter. All promoters have been divided into the three groups in accordance with the calculated distances: 1) GC-group (sequences are close to GC consensus and far from CT consensus); 2) CT-group (sequences are close to CT consensus and far from GC consensus); 3) Z group (sequences are far from both consensi). CT found to be the largest group. It includes 406 promoters. Z group contains 339 promoters and GC contains 93.

We have analysed these groups of promoters in order to understand whether our classification reflects functional, taxonomic or some other peculiarities of the examined genes. We have investigated possible correlation between these classifications and the key words associated with these promoters in the EPD database. The most interesting examples of the revealed correlation are shown in Table 2. Statistical significance of the obtained correlations was calculated on the basis of binomial distribution.

Table 2.Examples of revealed correlation between obtained classification of promoters and their functional, taxonomic and other characteristics.

Group of promoters	Keywords	N [1)]	k [2)]	Probability [3)]
GC	1.1	118	99	1.1×10^{-15}
GC	1.1 + 6.2	130	110	0.8×10^{-19}
GC	1.1 + 3.3	125	106	2.2×10^{-17}
CT + Z	Hs + Rn	240	166	1.1×10^{-8}
CT + Z	6.1.5 + 6.1.4	196	144	1.6×10^{-10}
CT	6.1.2.5.1 + 6.1.4.2	16	11	0.8×10^{-9}

[1)] Number of promoters matching the keywords;
[2)] Number of promoters from N matching the consensus.
[3)] Probability to get such k to N case by random chance

It appears that groups of promoters characterized by distinct transcription initiation site correspond to different functional and taxonomic groups of genes. It turned out that the majority of plant promoters (class 1.1 in EPD) belong to GC-group. This group also includes a number of promoters of plasmides (class 1.2), mobile elements (class 6.2), viral genes (class 3.3). Vertebrate promoters are distributed uniformly among all three groups.

However, in some cases promoters of functionally related genes have clear similarity in the structure of their transcription initiation sites. For example, promoters of the genes encoding enzymes, hormones, grows factors and other regulatory proteins (classes 6.1.4 and 6.1.5) appears to be classified mostly to the groups CT and Z. Promoters of the actine genes (class 6.1.2.5.1) and genes for enzymes of nucleotide metabolism (6.1.4.2) are almost entirely included in the CT-group.

Thus the kernel method developed in this paper allows to determine the local consensi of functional sites in DNA sequences and to classify these sites. The obtained weight matrixes can be used for searching potential transcription start sites in genomic sequences.

Acknowledgements

Different parts of this work were funded by the Siberian Branch of Russian Academy of Sciences and by Volkswagen-Stiftung (I/75941).

References

1. Aivazyan S.A. Buhshtaber V.M., Enukov I.S., Meshalkin L.D. (1989) Applied mathematics. Classification and reduction of dimensions, Moskwa, "Finansi i Statistika", 608 pp.
2. Rogozin I.B., Milanezi L., (1997) Analysis of donor splice sites in different eukaryotic organisms. *J.Mol.Evol*, **45**, 50-59.
3. Kudo M., Kitamura-Abe S., Shimbo M., Iida Y (1992) Analysis of context of 5'-splice site sequences in mammalian pre-mRNA by subclass method. *Comput. Appl. Biosci*, **8**, 367-376.
4. Kondrakhin Y.V., A.E.Kel , N.A.Kolchanov, A.G.Romashchenko, L.Milanesi (1995) Eukaryotic promoter recognition by binding sites for transcription factors. *Comput. Applic. Biosci.* **11**, 477-488.
5. Kel A.E., Kondrakhin Yu.V., Kolpakov Ph.A., Kel O.V., Romashenko A.G., Wingender E., Milanesi L., Kolchanov N.A. (1995) Computer tool FUNSITE for analysis of eukaryotic regulatory genomic sequences. In: *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology,* AAAIPress, California, pp.197-205
6. Durbin R., Eddy S.R., Krogh A., Mitchison G. (1998) Biological sequence analysis. Probabilistic models of proteins and nucleic acids. *Cambridge University Press*, 346 p.

ANALYSIS OF THE REGION OF INTRON 6 OF THE HUMAN TDO2 GENE IN THAT POINT MUTATIONS ASSOCIATED WITH PSYCHIATRIC DISORDERS ARE LOCATED WITH THE AID OF COMPUTER AND EXPERIMENTAL APPROACHES

Merkulova T.I., Vasiliev G.V., Ponomarenko M.P., Kobzev V.F., Podkolodnaya O.A., Ponomarenko Yu.V., Kolchanov N.A.

Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: merk@cgi.nsk.su
*Corresponding author

Keywords: tryptophan oxygenase gene, binding site, YY-1 transcription factor, single nucleotide polymorphisms, psychiatric disorders

Resume

Single base mutations G → A at position 663 and G → T at position 666 of intron 6 of the human tryptophan oxygenase gene (*TDO2)* are associated with a variety of psychiatric disorders [Comings D.E. et al., 1996]. Binding of rat liver nuclear extract proteins to synthetic double-strand oligonucleotides corresponding to three allelic states of the region between 651 bp and 680 bp of human *TDO2* intron 6 has been studied by gel shift assay. It has been demonstrated that to each allelic state of the region there corresponds a specific set of proteins that interacts with it. With the aid of computer method based on the averaging of oligonucleotide frequencies and using specific anti-YY-1 antibodies it has been shown that both mutations damage the YY-1 transcription factor binding site.

Introduction

Functional site recognition is one of the basic methods for computer-assisted analysis of nucleotide sequences, and rather a promising approach to many problems of experimental molecular biology. In particular, there is an urge to develop reliable methods for regulatory protein binding site recognition for analysis of known and a search for the unknown regulatory regions of genes, identification of the groups of coherently regulated genes, reconstruction of signal transduction pathways on the basis of differential display or cDNA micro-array data and to solve other problems. In this work we took advantage of a combined approach based on a computer-assisted search for transcription factor binding sites and experimental techniques for analysis of the polymorphic region in intron 6 of the human TDO2 gene.

Tryptophan 2,3-dioxygenase (TDO2, EC 1.13.11.11) is the rate-limiting enzyme in the oxidative degradation of tryptophan, the serotonin precursor, which therefore controls serotonin level in the body. Defects in serotonin metabolism and abnormal serotonin/tryptophan levels have been reported for many behavioural disorders. This suggests the *TDO2* gene as a potential candidate gene in psychiatric genetics [1].

At present studies of the nucleotide sequence polymorphism in the candidate genes are one of the main approaches to investigation of genetic predisposition to polygenic disorders. In *TDO2* gene single nucleotide polymorphisms showing a significant positive association with drug dependence, Tourette syndrome and attention deficit hyperactivity disorder were identified [Comings,D.E. et al., 1996]. These are two nucleotide substitutions located much central in intron 6 (G→A at position 663 bp and G→T at position 666 bp relative to start of the intron).Single nucleotide substitutions reported in introns are often associated with various diseases. However, the mechanisms by which such mutations could affect gene expression normally remain unclear, unless the mutations affect splicing signals. We have demonstrated that the region between 651 and 680 bp from the start of the intron 6 of the human *TDO2* gene, in which polymorphic sites associated with behavioural disorders are located, bind liver nuclear proteins, and that every allelic states the region can be in is accounted for a specific sets of proteins that interact with it. The disturbance in the binding of the transcription factor YY-1 is one of the effects of the mutations.

Results and Discussion

We used a mobility shift assay to see if nuclear factors bind to double-strand oligonucleotides that correspond to the different allelic states of the 651-680 bp region relative to start of intron 6 of the human *TDO2* gene (WT, M1 and M2 oligonucleotides, Fig.1a). The extract proteins formed 3 complexes with the WT oligonucleotide, which corresponds to the most frequent allele (Fig.1b). Both nucleotide substitutions found in humans [Comings,D.E. et al., 1996] caused a dramatic change in binding pattern. The substitution of G by A at position 663 bp. (the M1 oligonucleotide) caused the band corresponding to complex 3 to abate, if not disappear, abruptly and a group of less mobile bands to occur ("complex 4"). In the case of the G → T substitution at position 666 bp (the M2 oligonucleotide) the intensity of the indicated band had been decreasing not so strongly, but there also appeared new slowly migrating bands ("complex 5").

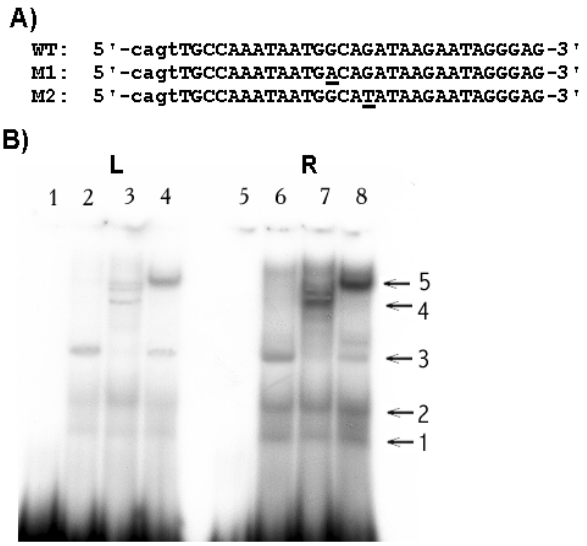


Figure 1. Distinct nuclear proteins bind to double-strand oligonucleotides comprising the sequences of three alleles of 651-680 bp region of intron 6 of the human *TDO2* gene. (**A**) Coding strands of oligonucleotides used: WT - most frequent allele, M1 and M2 - mutant variants. Point mutations are underlined, lower-case letters correspond to nucleotides added to make 5' overhangs. (**B**) Binding of nuclear proteins to WT (lanes 2, 6), M1 (3, 7), M2 (4, 8) oligonucleotides. Lanes 1, 5 - no extract. Oligonucleotides were incubated with 1 (**L**) or 4 (**R**) μg of rat liver nuclear extract protein.

To answer the question as to which proteins could bind to the WT, M1 and M2 oligonucleotides we did the search for regions homologous to the binding sites of various transcription factors using the TESS program package [2]. We found regions, similar to the binding sites for the GATA; HNF1 and Sp1 transcription factor families, and to the SRE and ANF elements. Double-strand oligonucleotides corresponding to such sites existing in real genes and previously used by other investigators for the same purposes, were synthesised and used as competitors in gel retardation experiments. The available oligonucleotides corresponding to the binding sites of the HNF3, C/EBP and AP1 families of transcription factors were taken as competitors.

As turns out, of the oligonucleotides used, only SRE (Fig.2) of the mouse *c-fos* gene promoter (-318 to -289) was competing against WT: in the presence of its 40-fold excess complex 3 would disappear. As is known, transcription factors YY-1, SRF and NF-IL6 bind to the mouse *c-fos* SRE [Treisman,R., 1992] which suggests that these proteins might be candidates for binding to the region in question.

A further theoretical analysis of the nucleotide sequences of the allelic states of the region in question was performed using a method developed for averaging nucleotide and dinucleotide frequencies at the positions of the site in question [Ponomarenko,M.P. et al., 1999]. The original data were 27 DNA fragments experimentally found to contain the transcription factor YY-1 binding site. The fragments were aligned using the standard method of multiple alignment maximising Gibbs potential [Lawrence C., 1994]. Then in every position of the DNA fragments being studied, the nucleotide and dinucleotide frequencies were determined, and a method was developed such that the YY-1 site is recognized by averaging the oligonucleotide frequencies. As can be seen from Fig. 3, a smaller second type error (overprediction) is thus produced than following the traditional method of nucleotide frequencies, irrespective of first type error (underprediction). This improved accuracy was due to the consideration of preferred "adjacent nucleotides". Using this method, an analysis of the nucleotide sequences of the allelic states of the region between 651bp and 680bp of human TDO2 intron 6 was performed. With each position tried for the start of the putative YY-1 site (position "0"of the site), the mean nucleotide and dinucleotide frequencies were calculated in all positions of the variant being analysed. The profile of similarity

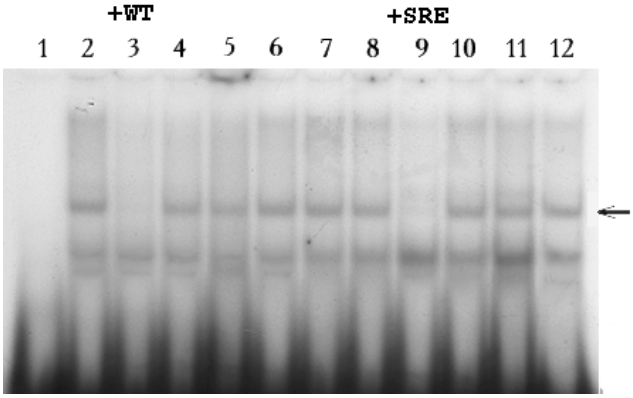


Figure 2. Competition of WT oligonucleotide-protein complex formation by various oligonucleotides. Lanes: 1, no extract; 2, no competitor;. 3-12, a 40-fold excess of unlabeled oligonucleotides: 3 - WT; 4 - M1; 5 - M2; 6 - AP1; 7 - HNF1; 8 - GATA; 9 - SRE; 10 - ANF; 11 - AP2; 12 - HNF3. Arrow denote the position of complex 3.

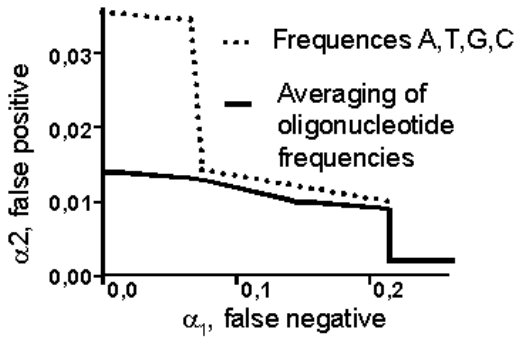


Figure 3. The method of YY-1 sites recognition by averaging of the oligonucleotide frequencies decreases type II error (overprediction) compared to traditionally used method of nucleotide frequencies

between YY-1 sites and WT is presented in Fig.4a, where dotted lines indicate the 95%-boundaries of similarity with YY-1 sites (upper) and random DNA (lower). As can be seen, the peak of similarity between WT and YY-1 sites at position 661 is over the upper 95%-boundary, which points (arrow) at region 661-667 bp of intron 6 of the human *TDO2* gene, which has a reliable similarity with the YY-1 sites. As turned out, the G→A (663) and G→T (666) mutations are located within the likely YY-1 site. In the case of G→A (663) mutation (variant M1), the similarity of the YY-1 sites with region 661-667 bp is lower than the lower 95% boundary, which implies that there is no YY-1 site in. If the mutation is G→T (666), the similarity of the YY-1 sites with variant M2 is to over the upper 95% boundary (Fig. 4b). As can be seen, quantitatively the pattern of similarity between the YY-1 and WT (strong), M1 (none) and M2 (weak) behaves as band 3 does while the liver extracts bind to these three variants (Fig.1b). As to the other two transcription factors, SRF and NF-IL6, which bind to the SRE-element of the *c-Fos* gene, [Treisman R., 1992], no binding site has been predicted for either in WT, M1 or M2 of 661-668 bp of intron 6 of the human *TDO2* gene by a similar analysis (Fig. 4b). This implies that among YY-1, SRF and NF-IL6, YY-1 is the most likely transcription factor to bind to region 651-680bp of intron 6 of the human *TDO2* gene.

Specific anti-YY-1 antibodies used in the band shift assay gave total support to this hypothesis. As is shown in Fig.5, addition of antibodies to the nuclear extract totally eliminates the protein-WT oligonucleotide complex represented by band 3, which is replaced by a slower migrating band. Complex 3 also disappears in the case of oligonucleotide M2, however no supershift can be seen, because this band is as mobile as one of the bands of the initial pattern.

YY-1 is a polyfunctional protein, which was shown to either repress or stimulate gene expression depending on the context of the corresponding regulatory region. As a component of the nuclear matrix YY-1 may also be involved in chromatin organisation possibly by tethering DNA to nuclear matrix. Whichever it is, it influences transcription intensity. Thus, both mutations damaging the YY-1 transcription factor binding site may result in a change in *TDO2* transcription level, which accounts for phenotypic changes.

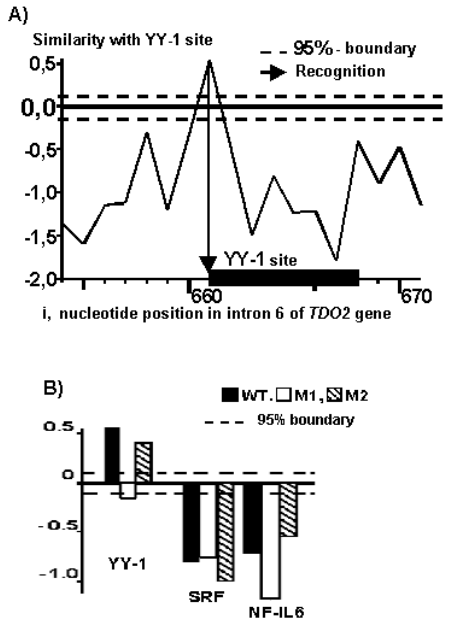


Figure 4. Recognition of the YY-1 transcription factor binding site in region 651-680 bp of intron 6 of human *TDO2.* **A)** The profile of similarity between YY-1 sites and WT as calculated by Formulae (2) and (3). Dotted lines indicate the 95%-boundaries of similarity with YY-1 sites (upper) and random DNA (lower). The peak of similarity at position 661,which is over the 95%-boundary, points out (arrow) the location of the potential YY-1 site in region 661-668 bp (black rectangle).

B) Similarity of WT, M1 and M2 with YY-1, SRF and NF-IL6 transcription factor binding sites.

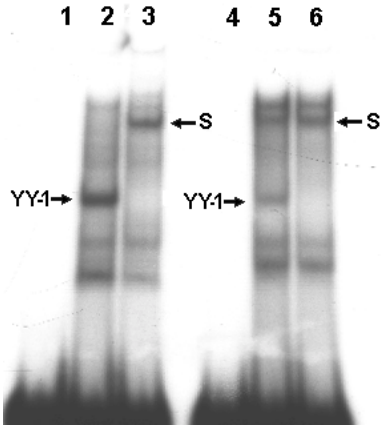


Figure 5. Effect of anti-YY-1 antibodies on mobility shifts of WT and MII oligonucleotides. Lanes 1, 4 - no extract. Lanes 2, 5 - 3 μg of nuclear extract, no antibodies. Lanes 3, 6 - 3 μg of nuclear extract was preincubated with 1 μl of specific anti-YY1 antibodies. The positions of **YY-1** containing complexes and supershift complexes (**S**)are indicated on the right.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (98-04-49654, 98-07-90126, 98-07-91078).

References

1. Comings,D.E., Gade,R., Muhleman,D., Chiu,C., Wu,S,, To,M., Spence,M., Dietz,G., Winn-Deen,E., Rosenthal,R.J., Lesieur,H.R., Rugle,L,, Sverd,J., Ferry,L., Johnson.J.P., MacMurray,J.P. (1996) Exon and intron variants in the human tryptophan 2,3-dioxygenase gene: potential association with Tourette syndrome, substance abuse and other disorders. *Pharmacogenetics* 6, 307-318.
2. Schug J., Overton G.C. (1997) TESS: Transcription element search software on the WWW. Technical report CBIL-TR-1997-1001-v0.0, of the Computational Biology and Informatics Laboratory, School of Medicine, University of Pensilvania.
3. Treisman R. (1992) The serum response element *TIBS*. 17, 423-426.
4. Ponomarenko,M.P., Ponomarenko,J.V., Frolov,A.F., Podkolodnaya,O.A., Vorobyev,D.G., Kolchanov,N.A., Overton,G.C. Oligonucleotide frequency matrices addressed to recognizing functional DNA sites (1999) *Bioinformatics*. 15, 631-643.
5. Lawrence C. (1994) Toward the unification of sequence and structural data for identification of structural and functional constraints. *Comput. Chem.*, 18, 255-258.

NON-CANONICAL SEQUENCE ELEMENTS AS ADDITIONAL SIGNALS IN PROMOTER RECOGNITION BY *E.COLI* RNA POLYMERASE

[1]Ozoline O.N., [2]Deev A.A., [3]Arkhipov I.V.

[1]Institute of Cell biophysics RAS, Pushchino, Russia
[2]Institute of theoretical and experimental biophysics RAS, Pushchino, Russia
[3]Moscow State University, Moscow, Russia
e-mail: ozoline@venus.iteb.serpukhov.su
*Corresponding author

Keywords: bacterial promoters, non-canonical elements

Resume

Motivation:

Among a variety of regulatory signals functioning in the chromosomal DNA promoters are the mostly studied and, probably, the least understood. Canonical sequences revealed by statistic analysis are the main promoter-specific determinants, which are common for all promoters recognized by vegetation form of RNA polymerase. These elements are composed of 12 base pairs (bp), however only 7 or 8 of them are usually found in the promoter DNA and approximately 10% of promoters have less than 7 canonical bp. That means that very low level of correspondence to the canonical elements may be sufficient for transcription initiation. However any combination of 4-8 canonical bp could be found in the genomic DNA of *E.coli* many times providing a possibility for faulty transcription initiation. Nonetheless exhausting production of false mRNA is not observed in the bacterial cell. That means that transcription machinery employs additional promoter-specific signals to distinguish promoter DNA from another sequences. Identification of additional promoter-specific elements as well as their detailed characterization is a long term objective of this study.

Results:

Non-canonical elements were recently revealed in the sequence of bacterial promoters. The significance in their presence was evaluated using non-promoter DNA fragments as well as sequence data of the *E.coli* genome. Elements were identified with statistical significance comparable to or even higher than for canonical hexamers. These elements are suggested as additional signals for promoter-search algorithms.

Availability:

The list of biochemically characterized promoter sequences with indicated non-canonical elements is available by Email: ozoline@venus.iteb.serpukhov.su.

Introduction

Newly revealed elements may be involved in the transcription initiation providing specific functional groups for interaction with RNA polymerase or creating specific structural features in the region of promoter DNA. In both cases they may be used by transcription machinery as an additional signals distinguishing promoter DNA from other sequences. The frequency in the presence of these elements in the genome of *E.coli* and in the promoter context was examined in this study with a purpose to estimate their significance relative to canonical hexamers TATAAT and TTGACA.

Methods and algorithms

To estimate the significance of the revealed elements we used the set of 441 experimentally characterized promoters and two control sets of non-promoter sequences (each containing 441 DNA fragments). One control set was composed of 110 bp long natural sequences taken from DNA of phages T7 and λ, while another set was combined of 110 bp long fragments taken from genome of E.coli. An importance of non-canonical elements as putative promoter-specific signals was evaluated by using the frequency in their presence in the DNA of E.coli. In both cases the level **B** of their significance was estimated employing non-parametric statistical method [Hollander,M. and Wolfe,D.A. 1973]:

$$B = (n-Nq)/(Nq(1-q))^{1/2},$$

where **n** is a number of promoters possessing certain sequence element in the region typical for its localization (indicated in Fig.1.); **N** is an overall promoter quantity (441); **q** is the frequency in the presence of this element in

the non-promoter fragments or in the genome of *E.coli* normalized to the length of the corresponding promoter region.

Results

Seventy six elements with well-established and putative functional significance were found by cluster analysis in the promoter DNA [Ozoline,O.N. et. al. 1997, Ozoline,O., and Deev,A., 1998] (Fig.1). Three different criteria were used to select these elements: a) all of them are dominant in one or another promoter region thus allowing promoter sub-grouping by clustering software; b) distribution of these elements along the promoter length showed a pronounced maximum in the region typical for their presence; c) quantity of promoters possessing these elements essentially (**B**>3) exceeds the value expected from their presence in the non-promoter DNA (Fig. 1). Taken into account that: a) analysis was performed for segments of different lengths and different levels of coincidence; b) clustering procedure is capable to reveal all patterns that occur imperfectly in the set of analyzed sequences; c) promoter-specific elements were searched without any prior assumption as to their quantity within one and the same promoter region, it is reasonable to believe, that the list of suggested elements is a redundant one for the region analyzed. Elements of the maximal length and their putative combination are presented in Fig. 1. Only some of them may be really involved in the transcription complex formation. Potential efficiency of these elements as a promoter-specific signals may be deduced on the basis of their distribution in the genomic DNA of *E.coli*.

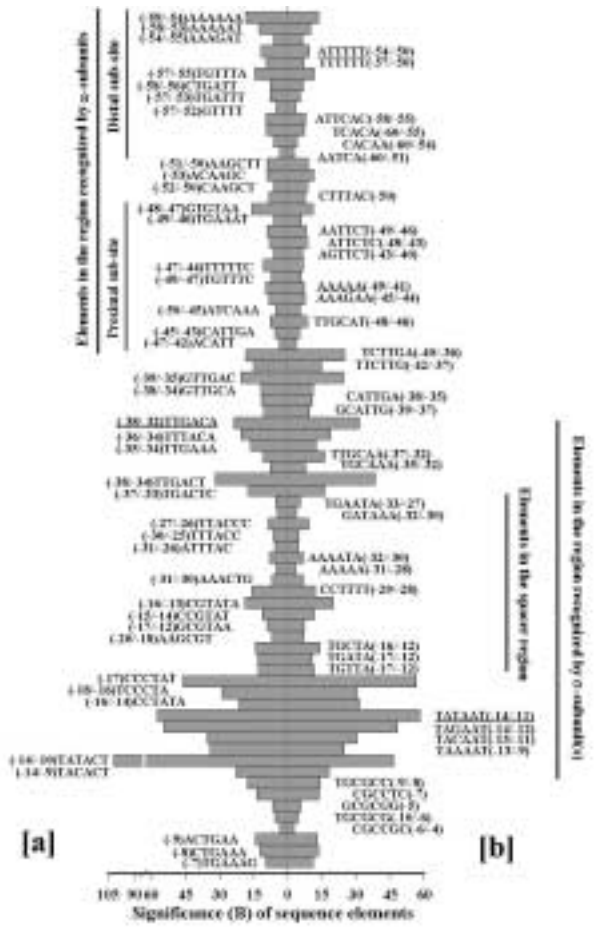


Figure 1. Statistical significance (**B**) for canonical (underlined) and non-canonical elements, calculated using the frequency in their presence either in non-promoter DNA (**a**) or in the genome of *E.coli* (**b**). Elements are sub-grouped on the basis of their sequence homology and preferred localization (indicated in parenthesis).

The value of parameter **B** calculated for canonical hexamers TATAAT and TTGACA is very high (57,7 and 23,5, respectively) what is due to the high frequency in the presence of these elements in the promoters and low frequency in their presence in non-promoter sequences (Fig. 1a). Surprisingly some motifs not identified by consensus analysis have the same statistical significance or even higher than for canonical elements. Thus, parameter **B** for TATACT and TTGACT, reaches the values 101,6 and 32,0, respectively. Several elements with very high statistical significance were found in the regions flanking canonical hexamers and in the upstream promoter areas. In the most cases the frequency in the presence of these elements in the non-promoter DNA is low, thus allowing a possibility of overestimating the parameter **B** even if 2 control sets of non-promoter DNA fragments were used to increase the evaluation reliability. That is why the statistical significance of newly revealed elements was reexamined using the frequency in their presence in the full genome of *E.coli* [Blattner,F.R. et.al. 1997]. Since promoter regions could not be strongly identified in the chromosomal DNA we compared the frequency in the presence of particular sequence motif in biochemically characterized promoters with the frequency of their presence in the whole genome. However the value of parameter **B** thus determined is underestimated, it displays a reliable lower limit for the statistical significance of any analyzed element. Fig. 1b represents the data obtained and practically reproduces the data of Fig. 1a.

Discussion

The data obtained clearly indicate that some elements in addition to the canonical hexamers may be used as markers of promoter regions. TATACT, TAGAAT and TATACT exhibiting very high statistical significance have the same localization as a canonical hexamers. Point mutations converting canonical hexamers to these sequences usually have no inhibitory effect. Three types of sequence elements are flanking –10 region at the 5'-end. One of them - dinucleotide TG at ~-15 is known as a target for interaction with σ-subunit. The highest statistical significance in this region is, however, displayed by CCCTAT, TCCCTA and CCTATA, providing a

possibility for their participation in the transcription complex formation. These elements, for example, may be recognized by alternative σ-factors (σ^{38} or σ^{32}), supporting their capability to activate vegetation promoters. Nucleotides TC and G frequently flanking 5'-end of the canonical element –35 can play the same functional role. Statistical significance of the elements typical for the upstream promoter region is less than of the non-canonical motifs found in the core promoter DNA. This region displays two distinct sites for interaction with two α-subunits. That is why motifs dominant in the distal sub-site usually resemble those in the promoter-proximal sub-site. However each of sub-sites could not be characterized by only one sequence element and sets of rather different motifs represent their structure. Some motifs may be a basis for the similar structural peculiarity recognized by α-subunits. For example, the geometry of the double helix formed by the AAAAA should have many commons with that of complementary TTTTT displaying similar features for interaction with RNA polymerase. The same possibility exists for dinucleotides TG and CA representing two other groups of non-canonical elements in the distal sub-site. At the same time, the structure formed by the former elements as well as configuration of their functional groups is essentially different from that of the later elements. That means that at least two different molecular mechanisms may be used by α-subunits for their docking on the promoter DNA and the possibility of alternative programs required for promoter identification should be taken into account in the improved promoter-search algorithms.

Acknowledgment

These studies are supported by the Russian Foundation for Basic Research (grant 00-04-48132).

References

1. Blattner, F.R., Plunkett III, G., et. al. (1997) The complete genome sequence of Escherichia coli K12. *Science*, **277**, 1453-1462.
2. Hollander, M., Wolfe, D.A. (1973) Non-parametric statistical methods. J. Willey and Sons, New York-London-Sydney-Toronto, pp. 15-26.
3. Ozoline,O.N., Deev,A.A. Arkhipova,M.V. (1997) Non-canonical sequence elements in the promoter structure.Cluster analysis of promoters recognized by *E.coli* RNA polymerase. *Nucleic Acids Research* 25, 4703-4709.
4. Ozoline,O., Deev,A.(1998) Non-canonical structural elements of promoter DNA and their role in interaction with RNA polymerase. *Molekularnaya biologia* 32, 441-446.

BIOCHEMICAL AND COMPUTATIONAL ANALYSIS OF TYPE I COLLAGEN GENE REGULATORY ELEMENTS

[1]Breindl M., [2]Mielke C., [3]Benham C.

[1]San Diego State University, San Diego, USA
[2]University of Aarhus, Denmark
[3]Mount Sinai School of Medicine, New York, USA
e-mail: mbreindl@sunstroke.sdsu.edu
*Corresponding author

Keywords: type I collagen, chromatin structure, distal regulatory elements, transgenic mice, enhancer, nuclear matrix, stress-induced DNA destabilization, computational analysis

Resume

The stage- and tissue-specific expression of many eukaryotic genes is regulated by *cis*-regulatory elements some of which are located in proximity to the start site of transcription whereas others are at considerable distances. In previous studies we have identified far upstream and downstream DNase I-hypersensitive sites in the murine alpha 1(I) collagen (*Col1a1*) gene which may play a role in the regulation of this abundantly expressed gene. Several of the 5' sites were tested for their biological functions in transgenic animals. One element located at -18 to -19.5 kb enhanced the position-independent activity of the linked *Col1a1* promoter and may be part of a locus control region. Another element located at -7 to -8 kb specifically enhanced reporter gene expression in the uteri of transgenic mice, suggesting that it contains a novel transcriptional enhancer that is involved in the steroid hormone dependent regulation of type I collagen expression in tissue remodeling in the uterus during the estrous cycle.

The molecular mechanisms by which distal *cis*-acting regulatory elements regulate transcriptional activity are poorly understood. The prevailing view is that the long-range chromatin structure and juxtaposition of proximal and distal regulatory elements are organized by the nuclear matrix, a highly branched intranuclear network of proteinaceous 10-nm filaments that is connected to the nuclear lamina. Interactions between chromatin and the nuclear matrix are mediated by biochemically defined scaffold or matrix attachment regions (S/MARs). These lack consensus sequences, but contain motifs that affect higher order DNA structure and can be identified by computational analysis as stress-induced DNA duplex destabilization (SIDD) sites. We found that several, but not all of the distal *Col1a1* regulatory elements show destabilized SIDD profiles as well as biochemical nuclear matrix binding activity, and that SIDD properties reliably predict S/MAR activity and *vice versa*. The proximal promoter also shows extensive SIDD potential and complex nuclear matrix interactions. These results support a chromatin loop model in which cooperation between distal and proximal regulatory elements is mediated by the nuclear matrix. The 5' ends of the murine *Col1a1* and the orthologous human gene revealed similar SIDD profiles, but limited DNA sequence similarity, suggesting that some DNA functions are evolutionarily conserved by preserving higher order DNA structural properties rather than nucleotide sequence.

Our results demonstrate the importance of DNA duplex destabilization properties, which are structural attributes not directly tied to the presence of consensus sequences or motifs, in the physiological functioning of DNA. They further document the usefulness of computational methods that predict SIDD properties in finding regulatory DNA sequences.

COMPUTER ANALYSIS REVEALS A SET OF ADDITIONAL PROMOTER ELEMENTS UPSTREAM OF MAIZE PLASTID GENES

Shahmuradov I.A., Akberova Y.Yu., Mustafayev N.Sh., Abdulazimova A.U., Aliyev J.A.

Institute of Botany, Baku, Azerbaijan
e-mail: il_shah@baku.ab.az
*Corresponding author

Keywords: computer analysis, plastid genes, nuclear-encoded RNA polymerase promoter, plastid-encoded RNA polymerase/σ^{70}/σ^{54}/σ^{32} promoters, regulatory elements, Zea mays, Escherichia coli

Resume

Motivation:

Plastid genes of higher plants are transcribed by at least two plastid RNA polymerases: the plastid-encoded RNA polymerase (PEP) and the nuclear-encoded plastid RNA polymerase (NEP). At least, some of plastid genes seems are transcribed from both PEP and NEP promoters. It may be indication of existence of various, alternative PEP and/or NEP promoters for other plastid genes. Moreover, most of the known PEP core promoters are reminiscent of σ^{70}-type eubacterial promoters. But it is also known that there are other different σ^{54}- and σ^{32}-type eubacterial promoters. In this regard one may also be argue the existence of σ^{54}- and σ^{32}-type promoters some of plastid genes. Besides, it is quite possible that there are some other regulatory elements resembling those for nuclear genes or exclusively related to plastid genes. Therefore we decided to search for both additional PEP/NEP promoters and σ^{54}/σ^{32}-type eubacterial promoters, as well as other transcription regulatory elements upstream of plastid genes of *Zea mays*.

Results:

The analysis revealed a set of additional putative PEP (σ^{70}-, σ^{54}-and σ^{32}-type) and NEP promoters upstream most of plastid genes of maize. Interestingly, some plastid genes contain several PEP promoters of various types (σ^{70}, σ^{54}, σ^{32}). Moreover, other motifs resembling plant nuclear and plastid transcription regulatory elements upstream most of maize plastid genes were found. Motifs of high homology in promoter regions of plastid and *Escherichia coli* genes were also revealed. Data obtained suggest that promoter architecture of plastid genes is more complex and includes various canonic promoters and other regulatory elements.

Introduction

Regulation of plastid genes' expression follows both nuclear and bacterial mechanisms. Plastid genes of higher plants are transcribed by at least two plastid RNA polymerases: the eubacterial-type plastid-encoded RNA polymerase (PEP) and the phage-type nuclear-encoded plastid RNA polymerase (NEP) (for review see: Maliga, 1998; Weihe and Borner, 1999). To date three types of NEP promoters have been identified [Weihe and Borner, 1999]. The NEP and PEP promoters do not seem to share identical sequence motifs. Some of plastid genes seems are transcribed by both PEP and NEP promoters [Maliga, 1998; Weihe and Borner, 1999]. This intriguing feature of plastid genes rises a question on the possibility of existence of various, alternative PEP and/or NEP promoters for other plastid genes. Moreover, most of known PEP core promoters are reminiscent of -10/-35 σ^{70}-type eubacterial promoters [Maliga, 1998; Weihe and Borner, 1999]. But it is also known that there are other different σ^{54}- and σ^{32}-type eubacterial promoters also consisting of two boxes. Interestingly, it is known that the promoter recognition specificity to PEP is conferred by three various σ-like factors of 67, 52 and 29 kDa. They were assumed to be nuclear-encoded and this prediction was recently confirmed for the Arabidopsis homologues of eubacterial σ-factors [Maliga, 1998]. In this regard one may also be argue the existence of σ^{54}- and σ^{32}-type promoters some of plastid genes. Besides, it is quite possible that there are some other regulatory elements resembling those from nuclear/eubacterial genes or exclusively related to plastid genes. Such, sequences that resemble enhancer- or silencer-like elements were detected in the 5'-region of *rpoB* genes from Arabidopsis, tobacco and spinach [Inada et.al., 1997].

To test these suggestions, we performed computer search for putative additional PEP/NEP promoters and σ^{54}/σ^{32}-type eubacterial promoters, as well as other transcription regulatory elements of nuclear and eubacterial origin upstream (-1000 : +50) of 108 plastid genes/ORFs (with exception of tRNA genes) of *Z.mays* in a such way: (i) a search for additional PEP (σ^{70}/σ^{54}/σ^{32}-type) and NEP promoters, as well as motifs resembling known plant nuclear and plastid transcription regulatory elements; (ii) pairwise comparison of 5'-promoter regions of maize plastid and *E. coli* genes for revealing possible common motifs for them (besides of known canonic promoter elements). Below we report some of the results of this analysis.

Methods and algorithms

DNA sequences of upstream regions of plastid and *E. coli* genes were extracted from GenBank by special computer programs GETFMT and GETSEQ developed by Shahmuradov. A comparative analysis of DNA sequences was carried out by "BLAST" [Altschul et al., 1997] and "GOMOL" [Solovyev et al., 1985] computer packages. To search for putative transcription regulatory elements "NSITE" computer method developed by Shahmuradov and Solovyev (see: http://genomic.sanger.ac.uk/gf/gf.shtml) was used.

σ^{70}-, σ^{54}- and σ^{32}-type promoter sequences of *E. coli* genes were extracted from GenBank and RegulonDB /version 2.0/ [Salgado et al., 1999]. About 200 plastid PEP/NEP promoter sequences and other plant regulatory elements of nuclear and plastid origin collected from literature have been used in the analysis (at present database PlantSite of plant regulatory elements is under construction).

Results and Discussion

A set of additional putative PEP and NEP, as well as σ^{54}/σ^{32} promoter sequences up-stream of many plastid genes have been found. Some of these elements are shown in Fig. 1. To date NEP promoters for 26 various plastid genes have been experimentally detected [Weihe and Borner, 1999. Our analysis suggests that, at least, 43 other plastid genes contain potential NEP promoters. Moreover, it is known that photosynthetic genes have only PEP promoters. But our data show that these genes may be also transc-ribed by NEP. At the same time our analysis confirms experimental data on transcrip-tion of *rpoB* gene exclusively by NEP. Nevertheless, a set of additional NEP promoter sequences upstream of the *rpoB* gene have been found (Fig. 1). Judge by the literature data, most of non-photosynthetic genes have promoters for both RNA polymerases. Besides, alignment of the 5'-promoter regions of *clpP* genes from various species revealed the new very conservative motifs (Fig. 2). Interestingly, these motifs upstream some of other genes are also found (Fig. 1). At last, to our knowledge, it is without precedence for some (total 24) plastid genes that they contain putative σ^{54}/σ^{32} promoters (Fig. 1). A possible role of these new PEP promoters is under question. Moreover, com-parison of 5'-promoter regions of maize plastid and *E.coli* genes detected various motifs of high homology (Fig. 3). The analysis also revealed existence of various DNA sites in promoter regions of chloroplast genes resembling known regulatory elements of nuclear and chloroplast origin.

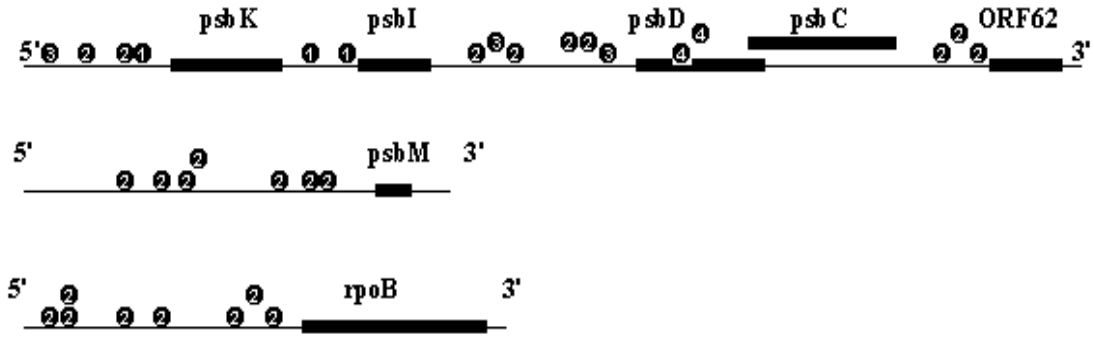


Figure 1. Localization of the putative promoters and new conservative promoter motifs of clpP genes (see: Fig. 2) in the 5'-promoter regions of photosynthetic psbM and psbK-psbI-psbD-psbC-ORF62 operon's genes, and rpoB gene coding a subunit of PEP. The positions of σ70/PEP (❶), NEP (❷) and σ54/σ32 (❹) promoters, and new promoter motifs (❸) are marked. The gene coding regions are indicated by more bold line.

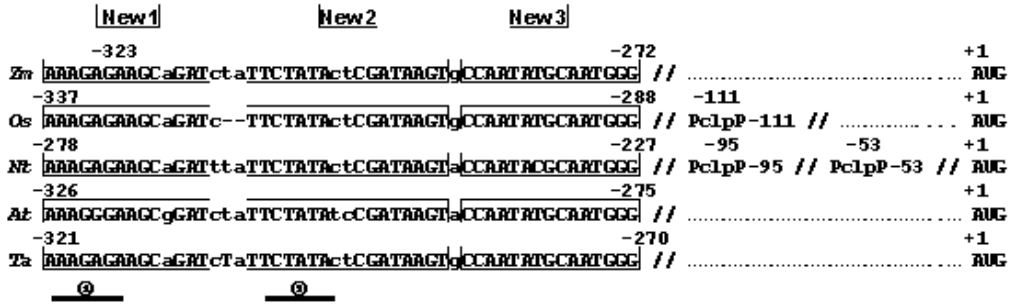


Figure 2. Partial sequence alignment of the maize (Zm), rice (Os), tobacco (Nt), Arabidopsis (At) and wheat (Ta) clpP promoter regions. The three conservative motifs (New 1, New 2 and New 3) are boxed. The native tobacco PEP promoter (PclpP-95) and NEP promoter (PclpP-53), and the rice NEP promoter (PclpP-111; Sriraman et al.,1998), as well as the first AUG codon are indicated. Motifs resembling the conservative boxes of the NEP promoter (Prps16-107) of the tobacco rps16 gene (Weihe, Borner, 1999) are underlined. Hereinafter positions of motifs are indicated in relation to the first AUG codon.

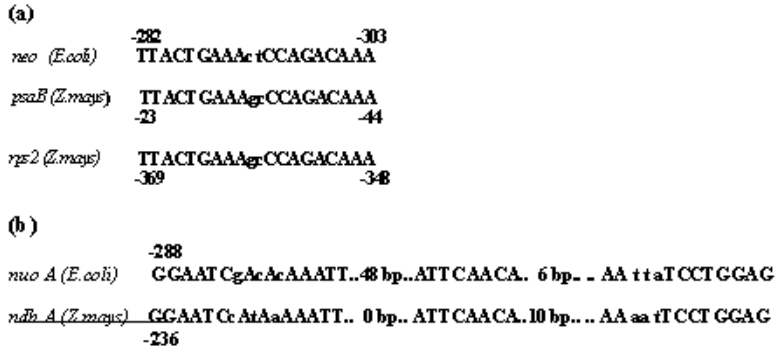


Figure 3. Homologous motifs (a, b) found upstream of E.coli and maize genes.

Thus, for example, a set of motifs similar (with one mismatch) to the regulatory site ATTCAAA found in the promoter region of the pea metallothionein (*MT*)-like gene $PsMT_A$ are present in the upstream regions of genes coding different subunits of NADH dehydrogenase (data not shown). The products of these plant *MT*-like genes could have a role in trace metal ion homeostasis [Fordham-Skelton et al., 1997]. Such motifs have been also found in the promoters some of other plastid genes analyzed by us. A possible role of these abundant motifs found in various plastid genes remains to be elucidated.

Data obtained suggest that organization of plastid encoded genes' promoters is very complex and, at least, some of plastid genes perhaps share functional regulatory elements similar to those in nuclear or bacterial genes.

References

1. Altschul, S.F., Madden, L.T., et.al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25, 3389-3402.
2. Fordham-Skelton A.P., Lilley C., Urwin P.E., Robinson N.J. (1997). GUS expression in *Arabidopsis* directed by 5' regions of the pea metallothionein gene $PsMT_A$. *Plant Mol.Biology*, 34, 659-667.
3. Inada, H., Seki M., et.al. (1997) Existence of three regulatory regions each containing a highly motif in the promoter of plastid-encoded RNA polymerase gene (*rpoB)*. *The Plant Journal*, 11, 883-890.
4. Maliga, P. (1998) Two plastid RNA polymerases of higher plants: an evolving story. *Trends in Plant Science*, 3, 4-6.
5. Salgado, H., Santos, A. et al. (1999) RegulonDB (version 2.0): a database on transcriptional regulation in *Escherichia coli*. *Nucleic Acids Res* 27, 59-60.
6. Solovyev, V.V.,Zharkikh, A.A., Kolchanov, N.A. (1985) A method of contextual analysis of the polynucleotide sequenses.1. The direct repeats within gene regions of b-, b'-, s-subunits of *Escherichia coli* RNA polymerase. Mol.Biol., (Russ) 19,524-536
7. Sriraman, P., Silhavy, D., and Maliga, P. (1998) The phage-type PclpP-53 plastid pro-moter comprises sequences downstream of the transcription initiation site. *Nucleic Acids Res.*, 26, 4874-4879.
8. Weihe, A., and Borner, T. (1999) Transcription and the architecture of promoters in chloroplasts. *Trends in Plant Science*, 4, 169-170.

A/T-TRACES IN THE INITIALLY TRANSCRIBED REGIONS OF BACTERIAL PROMOTERS. PUTATIVE FUNCTIONAL SIGNIFICANCE

Chasov V.V., Masulis I.S., Ozoline O.N.

Institute of Cell Biophysics RAS, Pushchino, Russia
e-mail: vvchasov@mail.ru
*Corresponding author

Keywords: promoter, non-canonical elements, initially transcribed region, A/T-tracts, transcription elongation

Resume

Motivation:

Basic characteristics of nucleotide sequences that distinguish promoter regions from non-promoter DNA have been and remain the subject of intensive study. In addition to the elements functioning in the core promoter region controlling the frequency of transcription initiation some additional signals were recently identified in the structure of initially transcribed regions [Ozoline et.al., 1999]. This region has random distribution of any particular base pair, at the same time it could be characterized by the presence of A/T-traces at positions +6, +23, +40 and +56. These elements have very high statistical significance, however their functional importance is to be elucidated. Here we present the first experimental data on this subject.

Results:

As a first insight to the functional role of newly revealed elements we analyze a possibility of their involvement in the processivity of RNA synthesis. Single-round transcription driven from the promoter D of phage T7 was used as an experimental approach. This template contains A/T-traces near all preferred positions. Our data indicate that point mutations removing A/T-traces at position +41 or +57 are accompanied with appearance of the shortened RNA-products thus supporting suggested possibility and outlining the trend for the further studies.

Introduction

A growing amount of experimental data indicate that sequence peculiarities in the initially transcribed regions may be important for productive transcription participating in the process of promoter clearance or other regulatory mechanisms. Downstream sequences of the promoters recognized by RNA polymerase of E.coli have recently been compiled and statistically characterized by clustering procedure [Ozoline et.al., 1999]. This approach allowed us to reveal a non-random distribution of A/T-traces with preferred localization at four distinct positions in the transcribed area. The observed maxima follow a rather strong periodisity close to 17 bp and are phased with the distribution of A/T-traces in the structure of core promoter DNA. Systematic analysis of this newly revealed feature is required both to understand the functional significance of periodically distributed elements as well as to estimate their applicability for promoter-search algorithms.

Materials and Methods

DNA templates: 213 bp fragment containing native promoter T7D (from –131 to +82 according to the start point of transcription) was used as an initial template. Point mutations removing A/T traces at two different positions were introduced at position +42 (T $\rightarrow$ C) or +58 (A $\rightarrow$ G). All fragments were obtained as a PCR-products synthesized from T7 DNA using corresponding primers.

Single-round transcription was performed in vitro as described by Kajitani and Ishihama [1989] . Open complexes were formed in the standard reaction mixture [Masulis et.al., 1998]. Transcription was initiated by substrate addition in the presence of heparin, which excludes RNA polymerase recycling, a (50μl) aliquots were removed from the reaction mixture at the fixed time (indicated in the Fig. 2) and mixed with an equal volume of the stop solution [Kajitani and Ishihama, 1989]. The samples were heated for three minutes at 90^0C and loaded on the 0.4mm polyacrylamide gel containing 8M urea to separate RNA products of different length. Gels were exposed for autoradiography employing Retina x-ray films at -18^0C.

RNA sequencing. The length of the products was determined by using 3'-dATP and RNA chain termination method [Krohn and Wagner, 1996].

Results and Discussion

The functional significance of the revealed feature has to be elucidated by systematic step-by-step analysis. Two principle models were suggested as initial step. Both models take into account the high affinity of RNA polymerase for A/T-rich DNA. The first one suggests that the enzyme may use properly positioned A/T tracts to modulate certain tight contacts with a core promoter region during promoter clearance and initial RNA synthesis. In this case it is reasonable to expect some influence of these elements on the transcription processivity. Another model takes into account phasing of periodically distributed elements with A/T-traces within core promoter region and proposes that RNA polymerase might use them for initial doking at the promoter region in the orientation optimal specific interaction. In this case it is reasonable to expect that the enrichment of the promoter region with A/T-traces might affect association constant with RNA polymerase or even the final configuration of the transcription complex. The aim of this study is to test the possibility of the first model.

Promoter T7D with the native transcribed sequences up to +82 was used as initial template. It contains A/T tracts near all optimal positions: at +7; +20, +41, +57 (Fig. 1) and belongs to the large group of promoters (approximately 20%), which have at least three consecutive A/T base pairs in the vicinity (±3) of all critical positions in the transcribed region (+6, +23, +40, +56). Two point transitions ($T_{+42}\rightarrow C$ or $A_{+58}\rightarrow G$) were introduced to study the effect of A/T traces on the processivity of RNA synthesis (Fig. 1).

```
+
          +10        +20        +30        +40        +50
gTTGGCTTTA GGATGGACGT TAGGAGGTGA CTTTAGGAGG ATACTTTAGG  AGACTGTAAC
                                          DC42 ↓           DG58 ↓
                                            C                G
```

Figure 1. Nucleotide sequence in the initially transcribed region of promoter T7D. Start point of transcription is indicated by lower case letter. Point substitutions are indicated by arrows. A/T traces at preferred for their distribution positions are underlined.

Figure 2 shows the products of single round transcription driven by three template derivatives. The RNA of expected length (82) is a major product in all cases. Abortive products usually do not exceed the length of 12 bases, therefore transcripts ≤11 bases should be considered as abortive ones. Several additional gel bands correspond to shortened RNA-products synthesized in the course of productive elongation. An appearance of these products is conditioned by the sequence of the template used. Point substitution $T_{+42}\rightarrow C$ is a reason of appearance of new transcripts with a length of 44/45 basis and some longer having no effect on the synthesis of shorter products. Point substitution $A_{+58}\rightarrow G$ affects the synthesis of the products with length of 55/65 nucleotides. In both cases the main effect of the point mutations on the rate of transcription elongation is observed in their close vicinity. Thus the data obtained support the possibility that A/T traces in the initially transcribed region may affect the processivity of RNA synthesis. Taken into account this functional significance as well as the fact that the frequency in the presence of A/T-traces in their preferred positions in all cases exceed 5 standard deviations [Ozoline et. al. 1999], the pattern of periodically distributed A/T-traces in the transcription regulatory region could be recomended as a distinguishing parameter of new type.

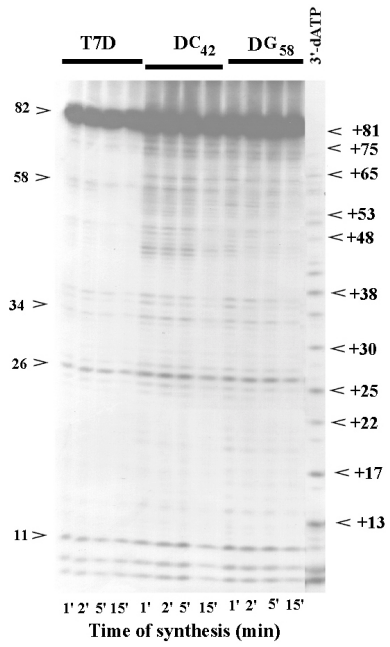


Figure 2. RNA products synthesized by RNA polymerase from T7D using native template or two mutant derivatives (DC_{42} and DG_{58}). Numbers below indicate transcription elongation times in minutes. The right lane contains RNA products obtained using RNA chain termination by 3'-dATP.

Acknowledgments

These studies are supported by the Russian Foundation for basic research (grant 00-04-48132) and Havemann scholarship within the frame of «Natural Scientists Initiative «Responsibility for Peace».

References

1. Kajitani, M., Ishihama, A. (1989) Promoter selectivity of Escherichia coli RNA polymerase. Differential stringent control of the multiple promoters from ribosome RNA and protein operons. *J. Biol. Chem.*, v. 259, p. 1951-1958.
2. Krohn, M. and Wagner, R. (1996) Transcriptional pausing of RNA polymerase in the presence of Guanosine tetraphosphate depends on the promoter and gene sequence. *J. Biol. Chem.*, v. 271, p. 23884-23894.
3. Masulis, I.S., Chasov, V.V., Kostyanicina, E.G. and Ozoline, O.N. (1998) Some aspects of protein-DNA recognition in transcription initiation. *Molekulyarnaya biologia*, v. 32, p. 598-602.
4. Ozoline O.N., Deev A.A., Arkhipova M.V.,.Chasov V.V., Travers A. (1999) Proximal transcribed regions of bacterial promoters have non-random distribution of A/T-tracts. *Nucleic Acids Res.* v. 27, p. 4768-4774.

CONSTRUCTION OF THE MODULE STRUCTURE MODEL OF THE REGULATORY SITE ON THE BASE OF THE MULTIPLE RELATIONSHIPS BETWEEN SITE POSITIONS

Kondrakhin Yu.V., Rogozin I.B., Romaschenko A.G.

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

Keywords: weight matrix method, site position relationships, the A and B boxes of tRNA genes, promoter, alignment, cluster analysis

Resume

A method for the recognition of the functional sites making use of multiple relationships between site positions was developed. In contrast to the canonical methods based on weight matrix and consensus, the new method considers not only independent nucleotide distributions at all site positions but also the significant interactions between different site positions. The proposed method was applied to the intragenic promoters composed by the A and B boxes in eukaryotic tRNA genes. Statistical treatment of nucleotide distributions at box positions demonstrated specific interactions between the positions within each box as well as between positions belonged to different boxes. Cluster analysis revealed the considerable heterogeneity within the A box. On this basis, the 12bp A1 and 11bp A2 subclasses were distinguished. All specific features of position interactions within the B box and each of the subclasses of the A box enhanced the efficiency of the promoter recognition in eukaryotic tRNA genes.

Introduction

The weight matrix method is one of the most useful in tool kits for describing and recognizing various functional sites. A large number of matrices representing the structure of diverse transcription factor binding sites has been accumulated in the TRANSFAC database [1]. Computer programs, such as MatInd and MatInspector [2], exploiting the whole set of the TRANSFAC matrices are now widely used for analysis of the regulatory regions of genes transcribed by polymerase II. GENSCAN [3], developed for theoretical analysis and prediction of the human genes, is another program efficiently using the matrix approach. Prediction of the donor and acceptor splice sites is an important capacity of the GENSCAN. It should be noted that search for donor motifs in the GENSCAN program is based on the matrix set thut takes the heterogeneity of entire known donor sites into account.

The highly repeated tRNA genes of eukaryots, interspersed among many chromosomes as single copies or clusters, are transcribed by polymerase III [4]. Initiation of transcription in all the diverse types of the tRNA genes is regulated by an intragenic promoter consisting of the far apart A and B boxes [5]. So far, consensi different in sizes for the A box and also in degeneracy at certain positions for the A and B boxes have been used to recognize them [4-7].

Although similar in promoter structure, the tRNA genes may have individual features that provide their differential expression. Therefore, a matrix embracing the widest possible range of site varieties appears to be more powerful than the consensus. The composition of the tRNA pool may vary, depending on the structure of the mRMA population in which codon frequencies are specifically distributed [8]. Furthermore, the idea underlying the consensus or matrix approach is independence of one site position from another. Even a cursory analysis of the composite elements [9] to which the tRNA promoters may be referred demonstrates relationships between particular positions of the A and B boxes [10]. These relationships are observed as interrelated substitutions within the sites. For this reason, recognition by the canonical methods may be not accurate enough because they leave unnoticed information concerning the potential interactions at certain site positions.

We suggest here a modified matrix approach whose recognition algorithm operates additionally with all the significant site position interactions revealed by analyzing a training sample of the tRNA gene sequences.

Materials and Methods

The tRNA gene sequences from the database available via Internet at http://www.uni-bayreuth.de/departments/biochemie/trna/ were utilized as training sample. These sequences served well the present purposes since their structure provided the needed information about the structural feature of the A and B boxes of the tRNA genes and, moreover, the presence of tRNA in cell cytoplasm indicated that their

promoters were functionally active. The type of the tRNA gene was identified by transition of anticodon in tRNA sequence into codon.

The basic information of the suggested method for recognition of either one box is provided by two parameter sets derived from aligned sequences of the training sample. The first set is a common weight matrix W=(wij), i=A,C,G,T, j=1,…,L, where L denotes the length of the site in question. Element wij of W matrix is the weight of the i-th nucleotide at the j-th site position. The second four-dimensional set of weights U=(uikjm), i,k=A,C,G,T, j,m∈ D, derived from analysis of dinucleotide distributions describing pairs of correlated site positions is novel in the theory of site recognition. The set D={(j,m): 1≤j,m≤L} includes only the site position pairs (j,m) whose relationships was found to be statistically significant.

The significance of the interaction of two different site positions was estimated by the chi-square test: the frequencies of the observed nucleotide combinations were compared with the expected in the case of position independence. The interaction range at positions i and m was estimated by the information content measure I(i,m)=E(i)+E(m)-E(i,m), where E(i) and E(m) denote the entropy of nucleotide distribution at position i and m, respectively, while E(i,m) stands for the entropy of mutual distribution of nucleotide pairs at both site positions. The higher are the values of information measure I(i,m), the stronger are the relationships between the positions. A I(i,m) measure approaching 0 means that there is no considerable dependence between the positions.

The site recognition procedure is centered on estimation of the score F(S) for the examined nucleotide sequence S=(s1,…,sL) from the formula

$$F(S)=\sum_{j=1,\ldots,L} w_{s_j,\,j} + \sum_{j,m\in D} u_{s_j,\,s_m,\,j,\,m}$$

The decision of whether a sequence S is a site or not is made by comparing the calculated F(S) score with the threshold obtained by the treatment of the training sample. It should be emphasized that function F estimates the similarity between the tested sequence S and the real sites in the training set used for calculating the weights wij and uikjm .

Results and discussion

Attempting to make the recognition method more efficient, we develop a procedure for aligning sequences composing the initial training site samples. The procedure differs from the canonical, which relies on the dynamic programming principle, in being iterative and more importantly, it focuses on identification not only of conserved site positions, but also of position relationships. Thus, the proposed alignment procedure is an important component of the recognition method aimed at obtaining the optimal parameters.

Furthermore, the novel alignment provides useful details about the structural features of sites. For example, modifications of the main alignment algorithm resulted in a procedure allowing an efficient cluster analysis for search of various site groups, with each having a common substructure. Analysis of the A box representing the eukaryotic Single Cell, Plant and Animal by the novel alignment procedure revealed different subclasses A1 and A2 of the A box. The different lengths of the A1 and A2 subclasses and the change of nucleotide distributions at certain positions resulted from a nucleotide deletion/insertion at position 9.

Table 1. Nucleotide frequency matrices calculated for the eukaryotic B box and the A1 and A2 subclasses of the A box.

a) Box B.

	1	2	3	4	5	6	7	8	9	10	11
T	26	1	594	644	3	1	0	228	403	6	110
A	105	4	50	1	1	149	645	284	58	0	26
G	506	640	1	0	1	495	0	93	4	1	6
C	8	0	0	0	640	0	0	40	180	638	503
	R	G	W	T	C	R	A	D	H	C	Y

b) Subclass A1 of the A box.

	1	2	3	4	5	6	7	8	9	10	11	12
T	246	6	0	11	196	26	1	6	209	135	5	45
A	8	201	7	1	6	13	253	37	21	10	1	7
G	1	46	241	2	33	0	0	211	14	34	248	200
C	0	2	7	241	30	216	1	1	11	56	1	3
	T	R	G	C	K	Y	A	R	T	B	G	G

c) Subclass A2 of the A box.

	1	2	3	4	5	6	7	8	9	10	11
T	389	2	11	197	35	191	0	302	8	8	2
A	0	78	12	0	40	11	91	7	0	0	1
G	2	303	355	4	141	190	298	11	376	386	387
C	0	8	13	190	175	80	2	71	7	7	1
	T	R	G	Y	S	B	A	R	Y	G	G

Table 1 presents the nucleotide frequency matrices for the B box and 2 subclasses of the A box. The frequency ratio of the A1 to A2 subclasses within the whole set of considered eukaryotic sequences is 39% : 61%. This is

in sharp contrast with the 74%: 26% ratio we obtained for DNA fragments homologous to the A1 and A2 subclasses in the whole set of prokaryotic tRNAs without the intragenic promoters that eukaryotes possess. Cluster analysis of the whole set of the nucleotide sequences of the B box failed to reveal any heterogeneity in it including the difference between eukaryotes and prokaryotes.

Upon further examination of the relationships between different positions within the same site, a set of interrelated positions became apparent. For example, the strongest relation was for positions 4 and 6 within the A2 subclass of the A box, I(4,6)=0.42. It is interesting to note that the relation between the same positions somewhat weakened during transition from eukaryotic to prokaryotic sequences, I(4,6)=0.30. However, the relation in this pair remained strongest compared to all the others. Table 2 compares nucleotide pair distribution at these positions in eukaryotes and prokaryotes It is of importance that positions close to each other and, as rule, alternating with conserved positions, forming distinct modules of a box, compose the most interrelated positions.

Table 2. Pairwise distribution at positions 4 and 6 within A2 subclass of the A box.

Nucleotide pair (n1,n2)*)	(T,T)	(C,G)	(C,C)	(C,A)	(T,A)	(T,C)	(G,T)	(C,T)
Pair frequency for eukaryotes	0.45	0.27	0.16	0.02	<0.01	0.04	<0.01	0.04
Pair frequency for eukaryotes	0.26	0.24	0.01	0.21	0.08	0.07	0.05	0.03

*) Nucleotide n1 is at 4-th position, nucleotide n2 is at 6-th position.

To enhance the efficiency of the recognition of the A and B box in the promoters of the tRNA genes, we additionally searched for interdependent positions in both boxes. The interactions were strongest for the A1 subclass at the following positions:

I(9,9)=0.18, I(10.8)=0.19, I(9,3)=0.11, I(6,1)=0.10,

I(5,1)=0.09, I(10,11)=0.08, I(8,9)=0.06,

and for the A2 subclass at the following positions:

I(6.8)=0.16, I(4,8)=0.12, I(5,9)=0.12, I(6,3)=0.11,

I(6.9)=0.08 I(8,8)=0.08.

Three methods for recognition of the A and B boxes – the canonical, making use of consensus and weight matrix, the novel, using position relationships – were compared. The novel method was more accurate. Thus, 93% of the examined boxes were identified in the individual A1 and A2 subclasses with the false positive rate 5-7 fold lower than for 2 canonical methods.

It is also of importance that the strong relations between the non-conserved positions within a box module makes module subdivision feasible. In general, a module organization is a characteristic feature of the eukaryotic and prokaryotic tRNA sequences. However, the module interaction between the A and B boxes was observed for only the eukaryotic tRNAs.

References

1. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 24(1), 238-241.
2. Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Res 23(23), 4878-4884.
3. Burge,C. and Karlin,S. (1997) Prediction of the Complete Gene Structure in Human Genomic DNA. J.Mol. Biol., 268, 78-94
4. Sharp,S.J., Schaack,L., Cooley,L., Burke,D.J. and Soll,D (1985) Structure and transcription of eukaryotic tRNA genes. CRC Crit. Rev. Biochem., 19, 107-144
5. Geiduschek,E.P. and Tocchini-Valentini (1988) Transcription by RNA polymerase III. Ann. Rev. Biochem., 57, 873-914.
6. Chalker,D.L. and Sandmeyer,S.B. (1993) Sites of RNA polymerase III transcription initiation and Ty3 integration at the U6 gene are positioned by the TATA box. Proc. Natl. Acad. Sci. USA, v.90, 4927-4931
7. Soazeiro,C.A.P., Kassavetis,G.A. and Geiduschek,E.P. (1996) Alternative outcomes in assambly of promoter compexes: the roles of TBP and a flexible linker in placing TFIIID on tRNA genes. Genes Dev., 10, 725-739
8. Ikemura,T. (1981) Correlation between the abundence of Escherichia coli transfer RNAs and occurrence of the respective codons in its protein genes. J.Mol. Biol., 146, 1-21
9. Kel,O.,V., Romaschenko,A.G., Kel,AA.E. , Wengender,E. and Kolchanov, N.A (1995) A compilation of composite regulatory elements affecting gene transcription in vertebrates, Nucl. Acids Res., 23, 4097-4103
10. Rogozin,I.B., Kondrakhin Yu.V., Naykova T.M., Yudin N.S., Voevoda M.I. and RomaschenkoA.G. (2000) The module organization of the A and B boxes in the tRNA intragenic promoter Proc. of the 2-nd BGRS Conference, Novosibirsk, ICG, SD RAS

DETECTING PATTERNS OF STRUCTURE-FUNCTION ORGANIZATION OF REGULATORY GENOMIC SEQUENCES IN A FIRST ORDER LOGIC

[1]Vityaev E.E., *Podkolodny N.L., Vishnevsky O.V., Kosarev P.S., Ananko E.A., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A.

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[1]Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
e-mail: pnl@omzg.sscc.ru
*Corresponding author

Keywords: knowledge discovery in databases, data mining, gene expression regulation, recognition of promoters

Resume

Motivation:

The knowledge on regulatory functions of DNA, RNA, and proteins is of the utmost importance for solving a wide range of molecular biological, molecular genetic, biotechnological, and medical problems.

Knowledge discovery in databases and data mining (KDD&DM) is an actively developing direction. It is a multistage interactive process, including “sample” formation, data “purification”, and their preprocessing; selection of data representation method (visualization, projection, feature extraction, etc.); isolation of a priori pieces of knowledge (background knowledge, domain theory, etc.); selection of the problem for knowledge discovery; selection of the necessary algorithm; interpretation of the results obtained, and application of the knowledge discovered.

The methods and algorithms for data discovery are to automatically process the information from databases (samples) to find the patterns significant for predicting activities or searching for functional sites.

Results:

The methods for data analysis, knowledge discovery in databases, and data mining for studying structure–function patterns of DNA, RNA, and proteins was developed. It is proposed to use the Theory of Measurements to represent this data type in the first order logic language, thus, in a relational form.

These methods allow all the stages of iterative process of KDD&DM directed to analysis and enrichment of the knowledge stored in the electronic library GeneExpress to be performed. Instrumental approaches were developed, that allowing procedures not only for solving individual problems, but also classes of problems to be generated.

The approaches described were tested while solving various problems. For example, the following hypotheses were tested while searching for patterns of various prokaryotic promoter types described with a help of redundant oligonucleotide motifs

Introduction

The knowledge on regulatory functions of DNA, RNA, and proteins is of the utmost importance for solving a wide range of molecular biological, molecular genetic, biotechnological, and medical problems. The data and knowledge of this type are being accumulated in the electronic library GeneExpress under development at the IC&G SB RAS [Kolpakov et al., 2000; Kolchanov et al., 1999].

The system GeneExpress was designed to accumulate the experimental data and provide for navigation, search for information, data analysis, and investigation of patterns of gene expression regulation. It integrates a great number of distributed databases and knowledge bases on structure–function patterns of DNA, RNA, and proteins and on the basic molecular genetic processes wherein these objects are involved together with hundreds of programs for detecting the structure–function patterns of genetic macromolecules significant for their function, predicting levels of their specific activities, recognizing, and classifying.

Knowledge discovery in databases and data mining (KDD&DM) is an actively developing direction. It is a multistage interactive process, including “sample” formation, data “purification”, and their preprocessing; selection of data representation method (visualization, projection, feature extraction, etc.); isolation of a priori pieces of knowledge (background knowledge, domain theory, etc.); selection of the problem for knowledge discovery; selection of the necessary algorithm; interpretation of the results obtained, and application of the knowledge discovered [Kovalerchuk and Vityaev, 2000].

As applied to the electronic library GeneExpress, the methods and algorithms for data discovery are to automatically process the information from databases (samples) to find the patterns significant for predicting activities or searching for functional sites. The methodical knowledge thus obtained includes the procedure (program, script, etc.) for searching for or predicting activity of a particular type of functional sites, comprising description of its purpose, application conditions, format of input data, limitations on the input data; format of

output data, etc. These pieces of knowledge allow, in turn, the complex scenarios for searching for various functionally significant regions of the genomic sequences annotated and predicting their structure and function to be synthesized automatically.

Detecting patterns in first order logic

The KDD&DM methods working in the language of first order logic are called relational DM methods [Kovalerchuk and Vityaev, 2000]. The system Discovery, which we are using for detecting sets (associations) of characteristics significant for the function of regulatory genomic sequences (RGS) and discovering automatically the knowledge on RGS structure–function organization and recognition, is a relational DM method.

Although essential variation (basic dissimilarity) in the structure–function organization is typical of the RGS with different functions, they all have two types of characteristics essential for their specific functions and quantitative activity levels, namely, (1) obligatory and (2) facultative.

Obligatory characteristics provide for the basal activity level of a particular RGS type. They are equal for all the RGS of this type in both their location and number. Obligatory characteristics are indispensable for RGS to perform their specific functions while interacting with regulatory macromolecules and regulatory supramolecular complexes, such as RNA polymerase, spliceosome, ribosome, etc.

The facultative characteristics provide for modulation of RGS activity levels with respect to the basal level, determined by obligate characteristics. Within a particular RGS, facultative characteristics differ in both their number and relative positions. It is the unique set of facultative characteristics present in a particular RGS and their relative locations that determine the specific activity level of this particular RGS.

Associations of facultative characteristics, interacting with both one another and obligatory characteristics of functional sites, determine their specific features, that is, the particular function of an RGS and the value of its specific activity [Kolchanov and Lim, 1994].

In the general form, the information on RGS is represented as a set of relational tables and may be, therefore, represented with a set of relations in the first order logic language. In particular, the five following tables allow the information on locations of the signals contained in any RGS group to be represented in a sufficiently general form:

Table 1: *<RGS superclass><RGS class>*

This table can be used to specify the hierarchical RGS classification in a form of relation of a kind_of type.

Table 2: *<RGS class><RGS name><RGS characteristics>*

This table determines the attribution of a particular RGS to this or that class. Actually, the name of RGS class is only one of RGS characteristics that may be used for formulating hypotheses. The table provides for the possibility of including additional RGS characteristics. If the RGS classification is not specified, it can be formally constructed according to the RGS characteristics.

Table 3: *<Signal superclass><Signal class>*

This table specifies the classification of signals.

Table 4: *<Signal class><Signal name><Signal characteristics independent of RGS and signal position>*

This table specifies the attribution of a particular signal to certain class of signals. The possibility of specifying additional characteristics of the signal independent of the RGS and positions wherein it is located is also provided for.

Table 5: *<RGS name><Signal name><Signal position><Signal characteristics specific for this position and RGS>*

This table describes a particular sequence. The number of signals observed and their position changes with individual sequence.

Thus, in the case of eukaryotic promoters as RGS superclass, a group of promoters of genes expressed in a particular tissue or organ exemplifies the RGS class; the name of a particular promoter belonging to this RGS class as specified in the TRRD database, the RGS name; and transcription factor binding sites or other specific conformational or structural characteristics significant for the promoter function, the signal. Signal position is usually determined relative to the gene transcription start. Undoubtedly, this table set is open to addition of the information on the processes wherein these RGS are involved, regulatory factors, etc.

Another major data type used in databases is a numerical representation of characteristics. In this case, the objects are represented as sets of the values of a characteristic. It is proposed to use the Theory of Measurements to represent this data type in the first order logic language, thus, in a relational form [Kovalerchuk and Vityaev, 2000]. According to the Theory of Measurements [Krantz et al., 1971, 1989, 1990], numeric values of characteristics are determined by relations. Axiomatic representations have been found for many characteristics. Following the Theory of Measurements, it is demonstrated how most common methods of data representation—object–characteristics tables, ordering and similarity matrices, and multiple and pairwise comparisons—can be represented in the first order language [Kovalerchuk and Vityaev, 2000]. In addition, the first order language allows various structures (graphs, nets, algebraic structures, etc.), characteristics lacking numerical representation (partial orders, tolerances, preferences, etc.), and mixtures of various characteristics to be represented. This is most important for knowledge discovery from a set of linked databases.

It has been demonstrated [Vityaev, 1992] that the system Discovery is capable of detecting any regular patterns displaying maximal estimates of conditional probabilities in the first order language. Therefore, the system Discovery is capable of discovering knowledge in GeneExpress.

The system Discovery generates hypotheses in a form of a parametric family of formulas of the following type:

$$A1\&\ldots\&An \Rightarrow A0, \quad (1)$$

where A0, A1,…,An are logical expressions (including logical connectives AND, OR, and NOT; brackets, and arbitrary arithmetic expressions with parameters). Parameters may be represented by serial numbers of characteristics, ranges of their alterations, their particular values, parameters modifying a characteristic (causing its various transformations), etc. The system allows hypotheses to be exhausted using a particular strategy, representing a semantic probability deduction [Krantz et al., 1971, 1989, 1990]. Hypotheses are refined through either adding new conditions to the premise or making substitutions.

The approaches described were tested while solving various problems. For example, the following hypotheses were tested while searching for patterns of various prokaryotic promoter types described with a help of redundant oligonucleotide motifs:

$$\forall a \exists p1,p2,\ldots,pi((Pos(p1) < Pos(p2))\& (Pos(p2) < Pos(p3))\&\ldots\&(Pos(pi-1) < Pos(pi))\&$$
$$(Sign(p1) = s1)\& (Sign(p2) = s2)\&\ldots\& (Sign(pi) = si) \Rightarrow (Class(a) = clj)),$$

where a is the oligonucleotide sequence; p1, p2,…,pi, oligonucleotide numbers; Pos(pj), number of oligonucleotide pi position in the sequence a, j = 1,…,i; Sign(pj), sign of oligonucleotide; s1,…,si $\in$ {+,-}, direct (+) or inverse (–) order of oligonucleotide; and Class(a), number attributed to the class whereto sequence a belongs.

Running of the program allowed a great number of patterns to be discovered, in particular, the following:

$$\forall a \exists p1,p2,\ldots,p10 ((Pos(p1) < Pos(p2))\&(Pos(p2) < Pos(p3))\&\ldots\&(Pos(p9) < Pos(p10))\&(Sign(p1) = s1)\& (Sign(p2) = s2)\&\ldots\& (Sign(pi) = si) \Rightarrow (Class(a) = clj))$$

p1=9, p2=27, p3=3, p4=4, p5=2, p6=1, p7=3, p8=2, p9=11, p10=6;

s1= +, s2= -, s3= -, s4= -, s5= -, s6= -, s7= +, s8= +, s9= +, s10= +; Clj = 1;

This means that the promoters where oligonucleotides are located according to the order 9 < 27 < 3 < 4 < 2 < 1 < 3 < 2 < 11 < 6 and the corresponding order (direct and inverse) belong to class 1.

Another problem used to test these approaches is connected with recognition of functional groups of promoters, in particular, promoters of the genes involved in regulation of lipid metabolism, erythroid -specific genes, and interferon-inducible genes, described in the TRRD database.

Putative regulatory sites detected in promoters using weight matrix technique were used for promoter description in addition to the experimentally found regulatory sites, described in TRRD.

Promoter primary sequences were extracted from the EMBL database. Weight matrices determined in the Transfac database were used. Random sequence samples constructed using the SAMPLES system, a part of GeneExpress, were used as a NOT class.

Conclusion

We are applying the methods for data analysis, knowledge discovery in databases, and data mining described in this work to studying structure–function patterns of DNA, RNA, and proteins. These methods allow all the stages of iterative process of KDD&DM directed to analysis and enrichment of the knowledge stored in the electronic library GeneExpress to be performed. Moreover, these approaches are instrumental, that is, allowing procedures not only for solving individual problems, but also classes of problems to be generated.

Acknowledgements

This work was supported by the Integrative Project of the Siberian Branch of the Russian Academy of Sciences No. IG2000/65, Russian Foundation for Basic Research (grants Nos. 98-07-91078, 99-07-90203, and 00-07-90337), National Human Genome Program, and The Committee of Science and Technology of the Russian Federation.

References

1. Kolpakov F.A., Podkolodnyi N.L., Lavryushev S.V., Grigorovich D.A., Ponomarenko M.P., and Kolchanov N.A. (2000) Methods for integrating nonuniform informational resources on regulation of gene expression in the electronic library GeneExpress. // *Programmirovanie*, 3, (in press), (in Russia).
2. Kolchanov N.A., Ponomarenko M.P. at al. (1999) GeneExpress: an WWW-oriented integrator for databases and computer systems for studying the eukaryotic gene expression. *Biofizika*, 44, No. 5, pp. 837–841.
3. Kolchanov N.A. and Lim H.A. (1994) *Computer Analysis of Genetic Macromolecules.* World Scientific, p. 556.
4. Kovalerchuk B. and Vityaev E. (2000) *Data Mining in Finance: Advances in Relational and Hybrid Methods*. Kluwer Academic Publishers, p. 308
5. Krantz D.H., Luce R.D., Suppes P., and Tversky A. (1971, 1989, 1990) Foundations of Measurement, 1–3. NY, London: Acad. Press.
6. Vityaev E.E. (1992) Semantic approach to development of knowledge bases. Semantic probabilistic derivation of PROLOG programs optimal for recognition using probability data model. In: *Logic and Semantic Programming* (Comput. Syst. 146), Novosibirsk, pp. 19–49.

NUCLEOSOME CODE ANALYSIS BY ESTIMATING MARKOV DEPENDENCIES

Orlov Yu.L., Levitsky V.G.

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

Keywords: chromatin structure, nucleosome code, stochastic complexity, Markov models

Resume

Motivation:

Recognition and mapping of nucleosome binding sites hold much significance for eukaryotic genome annotation. However, the code of nucleosome packaging is strongly degenerous. As foundation for its recognition serves a determination of all statistically significant constituent contexts (oligonucleotides), which specify nucleosome code by periodical alternation.

Results:

By applying the method of stochastical complexity estimation, we have analyzed the samples of nucleotide sequences containing nucleosome binding sites, these samples being extracted from the "Samples" database (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/) and the database developed by E. Trifonov (ftp://ftp.ebi.ac.uk/pub/databases/nucleosomal_dna/). We have determined the set of contexts that are statistically significant (non-random) for nucleosome packaging. This set consists of short oligonucleotides, the value of which was previously estimated by some other methods.

Introduction

Eukaryotic genomes are supplied with a specific chromatin structure. This structure is organized hierarchically at several levels. At the first level, double-stranded DNA is packed around a nucleosome, or histone octamer. In some genome regions, DNA is more densely packed into chromatin structure, whereas in the others it is less compacted. Notably, during the cell functioning and the organism's development, the extent of packaging varies. Those genome regions, where transcription is active, are more easily accessible and free out of chromatin relatively to regions at "silent" state. With this respect, determining of gene transcription activity could be based on information about levels of nucleosome packaging of these gene nucleotide sequences. However, significant information about such packaging could not be achieved simply out of the sequence content, because of extreme degeneracy of nucleosomal code [Trifonov E., 1997].

The nucleotide sequences, which are more preferred for nucleosome positioning, can be revealed by statistical methods [Widlund H., 1997]. In order to detect a nucleosome code, the search for periodicity of di- or three-nucleotide occurrences was performed [Satchwell S. et al., 1986; Ioshikhes I. et al., 1996; Stein A. & Bina M., 1999]. Besides, the DNA conformational features, significant from the point of view of molecular biology, were also considered [Levitsky V. et al, 1999]. Previously, the ascertainment of such oligonucleotides and estimation of significance of this ascertainment was made by different methods in various statistical models, but without accounting Markov properties.

Methods and algorithms

For recognition of nucleosome sites, a method could be applied, which is based on discriminant analysis by frequencies of oligonucleotides with different length [Levitsky V. et al., 2000]. However, an ascertainment of the length of oligonucleotides, the frequencies of which are accounted, should be made by empirical methods.

For analysis of DNA sequencies containing nucleosome positioning sites, we have used the method of significant contexts (oligonucleotides) ascertainment based on construction of a Markov model for generating the texts with unfixed order of Markov chain [Orlov Yu. and Potapov V., 2000].

We have considered a statistical model such that probability of the next-in-turn symbol occurrence in a communication is determined by preceding context, which defines the state of Markov chain and distribution of probabilities of the next occurring symbol (the set of numbers determining probabilities of A,T,G, or C occurrences after the context given). So, dependencies like $X_n...X_2X_1Y$ are taken into consideration, where Y is the symbol considered, X_i – preceding symbols, $1<i<n$, $n=5$ (the maximal length of the context n did not exceed 5 in our studies).

Contexts may be of various length, moreover, neither of them is the ending of another. It is convenient to make graphical representation of such contexts in a form of a tree. For DNA sequences, the tree will have 4 branches at each level, these branches corresponding to preceding symbols (Fig. 1). One should "read" the contexts in a tree according to the routes, which connect the leaves (suspended vertexes) with the root.

Our algorithm is founded on the method suggested by J. Rissanen [Rissanen J., 1983] and implemented for the tasks of data compression. This algorithm enables (i) to construct statistical model generating the sequence (generating source-tree with the contexts of interest) by the sequence content and (ii) to calculate the data complexity for this model. Automated determination of statistically significant contexts is based on evaluation of data complexity in the model together with complexity appearing due to adding novel parameters into the model [Orlov Yu. & Potapov V., 2000]. Thus, the problem of excessive parameters is being solved and the data are described more precisely.

We have used the sample of DNA sequences containing experimentally determined regions of nucleosome positioning [Ioshikhes I. & Trifonov E., 1993], i.e., 171 sequences with the length of 400 bp, along with the sets of DNA regions from the "Samples" database (http://wwwmgs.bionet.nsc.ru/cgi-bin/mgs/nsamples/), these sequences being classified by the organisms and by ability to bind to a nucleosome. Except dependencies of the form like $X_n...X_2X_1Y$, such that nucleotides follow each other directly, we have considered dependencies like $X_n\ [N]_k\ X_{n-1}[N]_k...X_2[N]_k\ X_1\ Y$, where N is an arbitrary nucleotide (generalized contexts). Insertions of arbitrary nucleotides between the significant ones have the length k that numbers from 1 to 20. Thus, setting more complicate Markov model has made an analysis of nucleosome code periodicity.

Implementation and results

The results of analysis of the sequences containing nucleosome binding sites and having the length of 400 bp are given in Fig. 1.

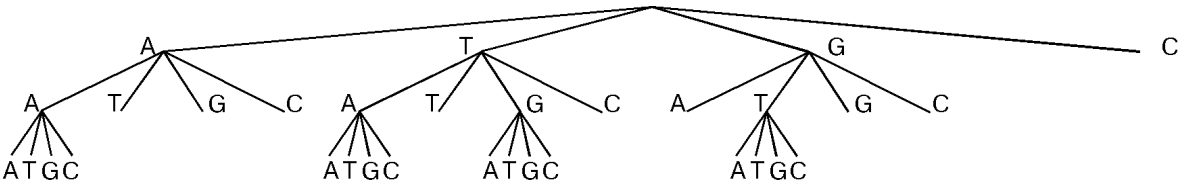


Figure 1. Context tree for DNA sequences with the length of 400 bp containing nucleosome binding sites.

Thus, the model generating the letters in DNA sequences containing the site binding with a nucleosome is described by the set of contexts with the length of either 3 nucleotides (NAA, NAT, NGT, NTG), or 2 nucleotides (TA, GA, CA, TT, CT, AG, GG, CG), or 1 nucleotide (C).

Here we set N={A,T,G,C}. Note that by the algorithm for construction of generating tree, the vertexes are directed downwards by 4 in number, because the length of preceding context either increases by a unit or stays constant. Nevertheless, such tree gives information about statistical significance of oligonucleotides and enables to indicate significant contexts. Interestingly, AA and TT dinucleotides providing the stiffness of double-stranded DNA were found within the set of significant contexts [Trifonov E., 1997].

The sequences of 400 bp in length have nucleosome binding site with the length of about 140 bp and the flanking regions, which may also influence statistical estimates. We have also considered the same sequences, but with the lesser flanking regions. The context tree for the sequences of 200 bp in length is illustrated in Fig. 2; for the sequences of 140 bp – in Fig. 3.

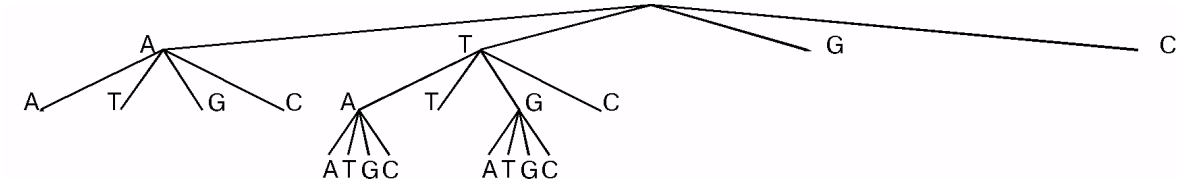


Figure 2. Context tree for DNA sequences, of 200 bp in length, which contain nucleosome binding sites.

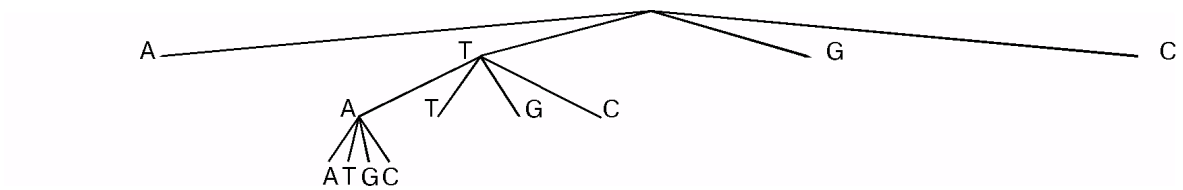


Figure 3. Context tree for DNA sequences, of 140 bp in length, containing nucleosome binding sites. The same tree of significant contexts was obtained for the sequences, complementary to analyzed sequences but disposed in inverse mode, from 3'- to 5'-end.

One may note that the tree constructed for the sequences with lesser lengths is incorporated into that for longer sequences.

By varying the length of a sequence with nucleosome binding site, from 400 to 140 bp, the tree with significant contexts has been changed and the number of branches diminishes to the state illustrated in Fig. 3 (data not shown).

Since DNA binds to a nucleosome in a double-stranded condition, we may suppose that the sequences complementary to those accumulated in databases are also capable to bind a nucleosome.

We have constructed the set of sequences that are complementary and inverse to the sequences with the length of 400 bp from the set analyzed previously. We achieved some set of context dependencies that do not correspond to Fig. 1. However, for the sequences containing only binding sites with the length of 140 bp, the tree of context dependencies was the same (see Fig. 3). This conclusion supports the evidence that two halves of DNA "coiled" over a nucleosome are symmetrical and that strict orientation of DNA sequence in the process of packaging is not necessary.

By considering context dependencies with insertions, of the type like $X_n\ [N]_k\ X_{n-1}[N]_k...X_2[N]_k\ X_1\ Y$, we can construct an analogous tree of context dependencies. Preliminary analysis has indicated that for real genetic sequences, there exist Markov dependencies of nucleotide frequencies that depend not only from the usual contexts, but from the generalized ones too. As usually, the order of the Markov chain n does not exceed 3. If the number k (number of inserted nucleotides between dependable positions) falls, the number of branches in a tree decreases, that is, the more remote dependencies are weaker.

We have generated random sequences of the same length with nucleotide frequencies equaling to those in real sequences. In this case, the context tree consists out of a single root. By generating random sequences with conservation of dinucleotides frequencies, the context tree has only a single level, i.e., a dependency only from one preceding nucleotide is observed. No long branches (dependencies upon two or more preceding symbols) were found. Thus, the context tree is capable to detect the complete set of significant preceding contexts, whereas the usage of only nucleotide frequencies does not entirely reflect statistical pattern of the sequences analyzed.

Statistical analysis of generalized contexts with insertions between significant symbols has proved that they should be also taken into account. For nucleosome code, we have noticed a periodicity of the signal by 10 bp, this periodicity being related to the length of double-stranded DNA spiral, which coils around histone octamer and contacts with it in particular regions (Fig. 4).

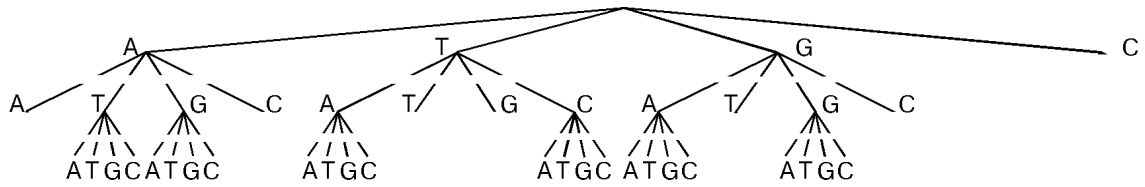


Figure 4. Statistical model for generating DNA sequences, which contain nucleosome binding sites, for generalized dependencies with periodicity equaling to 10 between the symbols $X_n\ [N]_k\ X_{n-1}[N]_k...X_2[N]_k\ X_1\ Y$, k=10. The tree is constructed for the set with the length of 400 bp.

Our method has revealed that there exist the dependencies both from preceding positions, and from positions, which alternate with periodicity 2,3,4,5, etc. (data is not shown). It is difficult to detect single particular period by the method considered, so, some other models are necessary for studying dependencies of positions in nucleosome binding sites. However, in generalized contexts, the contexts ending by A and T dominate not in all periods.

Discussion

Information accumulated in the publicly available databases [Ioshikhes I. and Trifonov E., 1993] on DNA sequences with experimentally detected nucleosome packaging (about 200 sequences, 10**4 bp) is too scarce in comparison with the data about the entire genome (10**9 bp). Experimental detection, whether this or those DNA sequence is involved into a nucleosome structure, is rather laborious and it does not provide the high level of accuracy of nucleosome signal mapping. Besides, complete understanding of gene functioning mechanisms is impossible without detection of their nucleosome packaging. Hence, it is of significance to detect by computer-assisted analysis what are the preferences of nucleotide sequence to bind to a nucleosome. For such computer analysis, it is necessary to use independent methods.

Markov model of generating the sequences binding to a nucleosome enables to consider the problem of nucleosome code from the novel point of view. Really, weak context code of nucleosome positioning [Trifonov E., 1997] makes us to suppose that the code is degenerate (that is, various DNA sequences are able to interact with histone octamer and to bind to a nucleosome), context signals are weak, and that the clear localization of a signal is absent. In Markov models generating symbol sequences [Durbin R., 1998], we do not use concrete signals or their localization inside the sequence, but bear upon dependency of the symbol occurrence upon

preceding context. Thus, as statistical model, Markov chain qualifies as description of nucleosome code and corresponds to theoretical standpoints about this code.

Coincidences in dependencies between direct sequence and inverted complementary one (i.e., likeness of the model context trees) indicates to the symmetry of hidden nucleosome code. In conclusion, the method developed makes possible to detect significant contexts, the data being in a good agreement with the previously known ones. In future, we plan to make detailed decoding of context features and to make biophysical interpretation of data obtained.

Acknowledgements

The authors are grateful to G. Orlova for help in translation of the manuscript into English, to N.A. Kolchanov, V. Potapov, O. Podkolodnaya, A. Katokhin and P. Kosarev for valuable comments and scientific discussion. The work was supported by Russian Foundation for Basic Research and Integration project SB RAS No 66.

References

1. Durbin R., Eddy S.R., Krogh A. and Mitchison G. (1998) *Biological sequence analysis: probabilistic models of protein and nucleic acids*. Cambridge University Press, 1-347.
2. Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., Trifonov E.N. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *J Mol Biol.*, **262**, 129-139.
3. Ioshikhes I., Trifonov E.N. (1993) Nucleosomal DNA sequence database. *Nucl. Acids. Res.*, **21**, 4857-4859.
4. Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Kolchanov N.A. (1999) Nucleosomal DNA property database. *Bioinformatics*, **15**, 582-592.
5. Levitsky V.G., Katokhin A.V., Kolchanov N.A. (2000) Inherent modular promoter structure and its application for recognition tools development. *Computational technologies (Novosibirsk)*, **5**, spec.issue, 41-48.
6. Orlov Yu.L. and Potapov V.N. (2000) Estimation of stochastic complexity of genetical texts. *Computational technologies (Novosibirsk)*, **5**, spec.issue, 5-15.
7. Rissanen J. (1983) A universal data compression system. *IEEE Trans.Inform.Theory*, **IT-29**, N.5, 656-664.
8. Satchwell S.C., Drew H.R. and Travers A.A. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol.,* **191**, 659-675.
9. Stein A., Bina M. (1999) A signal encoded in vertebrate DNA that influences nucleosome positioning and alignment. *Nucleic Acids Res.*, **27**, 848-853.
10. Trifonov E.N. (1997) Genetic level of DNA sequences is determined by superposition of many codes. *Mol Biol (Mosk)* **31**, 759-767.
11. Widlund H.R., Cao H., Simonsson S., Magnusson E., Simonsson T., Nielsen P.E., Kahn J.D., Crothers D.M., Kubista M. (1997) Identification and characterization of genomic nucleosome-positioning sequences. *J. Mol. Biol.,* **267**, 807-817.

CORRELATION ANALYSIS OF DNA CONFORMATIONAL CHARACTERISTICS OF HUMAN TOPOISOMERASE I CLEAVAGE SITES

Oshchepkov D.Yu., Kuzin F.E., Afonnikov D.A.

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

Keywords: DNA conformation, topoisomerase, cleavage site, correlation

Resume

Motivation:

Typical of topoisomerase I (topo I) cleavage sites are lack of consensus. Therefore, the conformational characteristics of the DNA helix might play the key role in determining the cleavage sites.

Results:

Correlation analysis of DNA conformational characteristics of human topo I cleavage sites has discovered the characteristics of double helix affecting presumably their recognition by the enzyme.

Introduction

DNA topoisomerases are enzymes controlling and maintaining the topology of DNA helix. Topoisomerase I is necessarily involved in eukaryotic transcription and replication (Stewart *et al.,* 1990; Annunziato, 1989). It is accepted that the enzyme relaxes the supercoils through cleavage–ligation of one DNA strand, thereby contributing to retention of particular DNA topology during these processes (Wang, 1996, Jupe *et al.,* 1995). It is assumed that the mechanism of topoisomerase I action is described by a scanning model of determining the cleavage sites. Despite a great volume of works on the mechanism of topoisomerase action, the cleavage site recognition, its particular properties, and the regulation mechanism are still vague (Caserta and di Mauro, 1996). Topoisomerase cleavage sites are, as a rule, essentially variable; therefore, no characteristic consensus is detected while their analysis. However, it was experimentally demonstrated that the sites of interaction with topoisomerase are not randomly distributed on DNA strand (Freeman and Garrard, 1992). Consequently, certain patterns of DNA sequence exist that provide for an increased affinity for topoisomerase I binding and consequent cleavage of certain DNA regions. The lack of consensus within these regions suggests that physico-chemical or conformational DNA characteristics underlie these patterns rather than contextual characteristics. Analysis of the cleavage sites using the system B-DNA-Video (Ponomarenko *et al.,* 1998) based on dinucleotide conformational characteristics has demonstrated that certain conformational characteristics of these sites are statistically significant for determining the topoisomerase cleavage.

In this work, we also used the method depending indirectly on contextual site characteristics. The analysis performed was based on pairwise correlations of dinucleotide conformational properties in DNA sequences. The results obtained suggested that the ability of topo I molecules to test the conformation of double helix within the interaction site underlies the regulation of this enzyme.

Materials

Results of two experiments on human topo I were selected for analysis, namely, (1) cleavage of human DNA restriction fragments by topo I from HeLa cells (Perez-Stable *et al.,* 1988) and (2) cleavage of SV40 DNA restriction fragments by topo I from 293 cells (Tsui *et al.,* 1989). Note that the conditions of these experiments (absence of camptothecin and low concentration of the enzyme) allowed a more precise selection and phasing of the sequence region to be performed. The samples formed are available in an electronic form with the database Samples (http://wwwmgs.bionet.nsc.ru/mgs/dbases/nsamples/). The sequences were phased with respect to the cleavage sites; the total length of the region analyzed equaled 80 nucleotides.

Method

Conformational DNA parameters calculated for nucleotide pairs (Ponomarenko *et al.,* 1999) were used for analyzing the aligned site sample.

The sample of N aligned (phased) sequences with the length L is considered. While analyzing, each sequence is recoded into a sequence of dinucleotides with the length $L–1$. The value of certain physico-chemical or

conformational characteristic ***f*** is juxtaposed to each dinucleotide. Thus, a matrix of ***N × L–1*** is formed. The element ***[l,m]*** of this matrix corresponds to the value of characteristic ***f*** of the dinucleotide located at position ***m*** of the sequence ***l.***

We took the value of correlation coefficient as the measure of dependence between the values of property ***f*** at different positions of the sample:

$r_{ij} = \frac{1}{N-1}\sum_{k}\frac{(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{d_{ii}\cdot d_{jj}}}$, where x_i is the property of the corresponding dinucleotide starting at position ***i***; ***N,*** number of sequences in the sample; and d_{ii}, sample dispersion of the characteristics of dinucleotide starting at position ***i.***

The correlation coefficient differing essentially from zero (that is, exceeding certain value $r_{critical}$ in its absolute value) indicates a statistical dependence between the values of the conformational property at the position pair ***i, j.*** The following equation is used for determining $r_{critical}$:

$t = |r_{ij}|\cdot\left(\frac{m}{1-r_{ij}^2}\right)^{1/2}$, which follows Student's distribution with ***m = N–2*** degrees of freedom (Anderson, 1958).

This method is based on conventional analysis of pairwise correlations in samples of amino acid sequences (Neher, 1994). However, analysis of nucleotide sequences lacks the problem of estimating the statistical significance within the sample due to evolutionary dependence of sequences (Afonnikov *et al.,* 2000), thereby simplifying the method.

Results and discussion

We studied the correlations of the values of 38 different conformational properties of dinucleotides (Ponomarenko *et al.,* 1999). Analysis of correlations of values of the mobility to bend towards the major groove, reflecting the capability of DNA helix to bend towards the major groove (Figs. 1, 2), gave most interesting results in both samples. A strong correlation between dinucleotide properties was found in the samples analyzed.

As is evident from Figs. 1 and 2, a great number of position pairs with significant correlation coefficients exists for this characteristic. A particular feature is in the fact that these pairs are organized in clusters up to 15 bp. In case of the first sample (Fig. 1), the clusters are mainly localized to 5' region relative to the cleavage site (clusters 1, 2, 3, and 5). In case of the second sample (Fig. 2), the positive dependence exists for the regions removed approximately 20 bp into 3' region and 30 bp into 5' region (clusters 1, 2, 3, and 5). In turn, the conformational nucleotide parameters in these clusters correlate negatively with characteristics of nucleotides located in the region around the cleavage site (clusters 4 and 6). These data indicate the existence of such regions where this nucleotide characteristic tends to deviate in a coordinated manner. Note an interesting effect—sometimes coordination of these deviations involves several regions. For example, the region corresponding to cluster 3 in the second sample is correlated with the regions corresponding to cluster 1 and 2, the properties within these regions deviating negatively and positively, respectively, as indicated by clusters located far from the main matrix diagonal (clusters 4, 5, and 6).

The analysis performed has demonstrated that the change in the direction of DNA helix mobility in the region of the cleavage site occurs in a dependent manner. It is likely that this correlation of the changes in this helix characteristic has a functional importance and a coordinated arising of stresses both in the site center and at its flanks might be important for determining the cleavage site. They may arise, first, in a torsionally strained DNA and, second, in the vicinity of two DNA segments crossing one another. Thus, we suggest that the supercoiling-induced conformational defect is the particular feature that determines the topoisomerase I site. This situation is clearer if we suggest a transition of torsion stresses into bending stresses by analogy with a torsionally strained tube made of a sufficiently soft metal (Fig. 3). Then, the bend arises within the region where the tube bendability is least even. If it is initially bent under the effect of external forces, the torsion stresses will increase the bend in this particular site—that is why topoisomerase I acts preferably in the regions of DNA strand crossings, where the strands undergo additional bending stresses.

Summing up, we suggest a correlation between the susceptibility to changes in DNA conformation and topoisomerase I action. In this situation, conformational characteristics of the site provide for (1) marking of the place where the enzyme interacts with DNA in case of topological hindrances, for example, in the vicinity of elongating transcription complex (Liu and Wang, 1987) and limit the topo I action to certain chromatin regions (Jupe *et al.,* 1995).

Acknowledgments

The authors are grateful to M.P. Ponomarenko for helpful advice, criticism, and ideas brought forth while discussing the work and to G. Chirikova for assistance in translation.

The work was supported by RFBR grants Nos. 98-07-91078 and 99-04-49879.

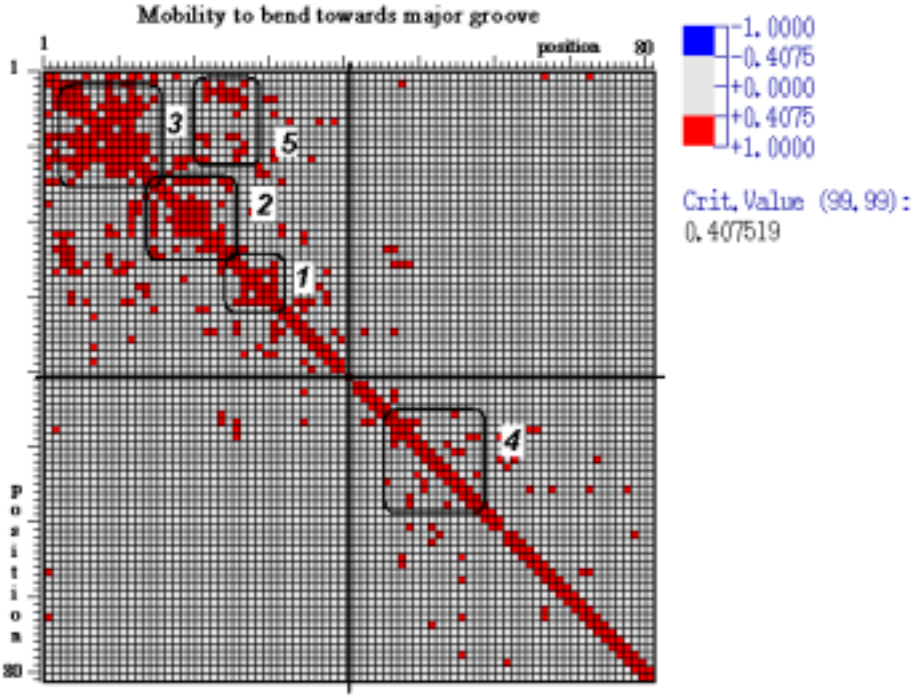


Figure 1. Matrix of correlation coefficients of the mobility to bend towards the major groove for a sample of 86 sequences (experiment 1) of DNA cleavage sites. Start positions of dinucleotides are indicated along the axes. Insignificant correlations are marked gray, significant are colored. The value scale is shown in the top right corner. The critical value is calculated according to equation (2) at a 99.99% significance level. The cleavage position is indicated with lines; clusters, with ovals.

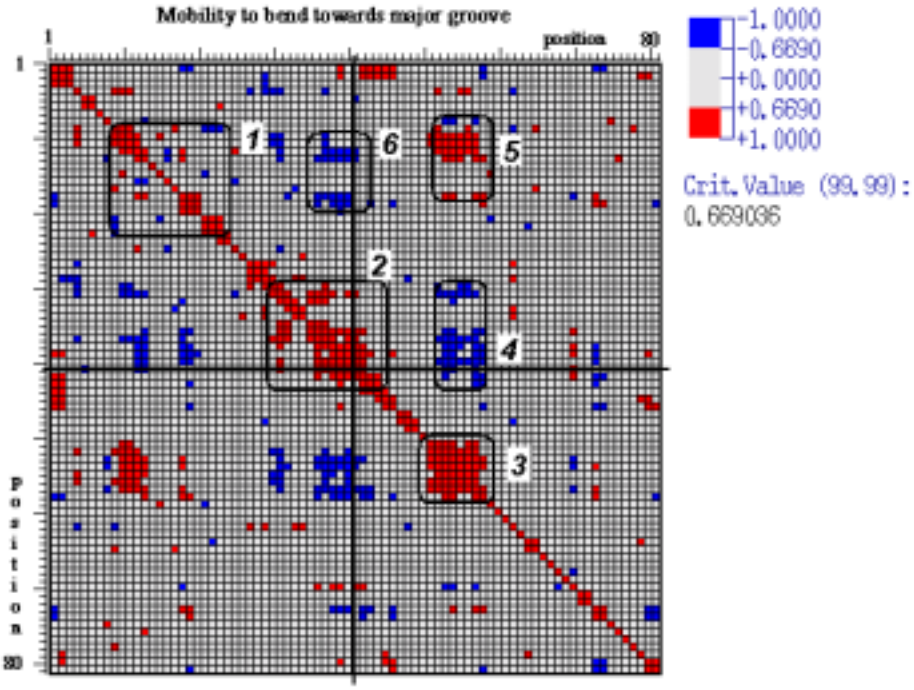


Figure 2. Matrix of correlation coefficients of the mobility to bend towards the major groove for a sample of 28 sequences (experiment 2) of DNA cleavage sites. See Fig. 1 for designations.

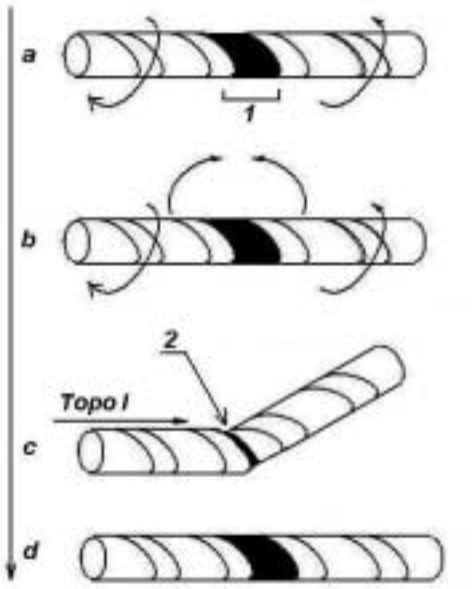


Figure 3. A possible mechanism of determining the topoisomerase I cleavage site (a) in the presence of torsional stresses and certain particular conformational characteristic of region (1); (b, c) forming the helix conformational defect and cleavage (2); and (c→d) stress relaxation through the site recognition and action of topoisomerase I (2).

References

1. Afonnikov D.A., Oshchepkov D.Yu., and Kolchanov N.A. Estimation of variances and covariances of protein physico-chemical characteristics in families of homologous sequences. *Comput. Technol.,* in press (2000).
2. Anderson T.W. (1958), An Introduction to Multivariate Statistical Analysis, John Wiley & Sons Inc., NY
3. Annunziato A. T., (1989) Inhibitors of topoisomerases I and II arrest DNA replication, but do not prevent nucleosome assembly in vivo. *J. Cell Sci.,* **93**, 593–603.
4. Caserta M. and di Mauro E. (1996) The active role of DNA as a chromatin organizer. *Bioessays,* **18**, 685–693.
5. Freeman L.A. and Garrard W.T. (1992) DNA supercoiling in chromatin structure and gene expression. *CRC Critical Rev. Euk. Gene Exp.,* **2**, 165–209.
6. Jupe E.R., Sinden R.R., and Cartwright I.L. (1995) Specialized chromatin structure domain boundary elements flanking a *Drosophila* heat shock gene locus are under torsional strain *in vivo*. *Biochem.,* **34**, 2628–2633.
7. Liu L.F., and Wang J.C. (1987) Supercoiling of the DNA template during transcription. *Proc. Natl. Acad. Sci. USA,* **84**, 1353–1358.
8. Neher E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci USA*, **91**, 98–102.
9. Perez-Stable C., Shen C.C., and Shen C-K. J., (1988) Enrichment and depletion of HeLa topoisomerase I recognition sites among specific types of DNA elements. *Nucleic Acids Res.,* **16**, 7975–7993.
10. Ponomarenko J.V., Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., and Kolchanov N.A. (1999) Conformational and physicochemical DNA features specific for transcription factor binding sites. *Bioinformatics,* **15**, 654–668.
11. Ponomarenko M.P., Vorobiev D.G., Ponomarenko J.V., Kuzin F.E., Gruzdev A.D., and Kolchanov N.A. (1998) Significant B-DNA conformational and physico-chemical properties of the DNA topoisomerase I sites. Proceedings of the First International Conference on Bioinformatics of Genome Regulation and Structure, (BGRS'98), ICG, Novosibirsk, **2**, 446–447.
12. Stewart A.F., Herrera R.E., and Nordheim A. (1990) Rapid induction of c-fos transcription reveals quantitative linkage of RNA polymerase II and DNA topoisomerase I enzyme activities. *Cell,* **60**, 141–149.
13. Tsui S., Anderson M. E., and Tegtmeyer P. (1989) Topoisomerase I sites cluster asymmetrically at the ends of the simian virus 40 core origin of replication. *J. Virol.,* **63**, 5175–5183.
14. Wang J.C. (1996) DNA topoisomerases. *Annu. Rev. Biochem.,* **65**, 635–692.

CORRELATION ANALYSIS OF HSF BINDING SITES CONFORMATIONAL PROPERTIES

*[1]Oshchepkov D.Yu., [1]Stepanenko I.L., [1]Afonnikov D.A., [2]Schroeder H.C.

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2]University of Mainz, Institute for Physiological Chemistry, Mainz, Germany
e-mail: diman@bionet.nsc.ru
*Corresponding author

Keywords: DNA, transcription factor, binding site, correlations, conformational properties, heat shock

Resume

Results:

A correlation analysis of DNA conformational properties for heat shock transcription factor (HSF) binding sites was made. These properties were calculated for nucleotide pairs. One of DNA helix properties, persistent length, was presumably found to influence HSF binding sites functioning.

Introduction

Different stress factors launch the synthesis of heat shock proteins (HSP), which are molecular chaperones defending the cell from damaging. Heat shock protein genes could be found in every cell type of each organism, from bacteria till humans. Due to stress impact, wrongly packaged and damaged proteins cause trimerization of heat shock factor (HSF), which activates HSP gene transcription by binding to promoters of these genes. HSF factor recognizes HSE sites represented by alternative inverted repeats of the nGAAn motif, where n is an arbitrary nucleotide. As a rule, HSP gene contains from two to four HSE sites. DNA-binding domain of every HSF forming a trimer interacts with one of GAA repeats. The typical site has 2-6 such repeats, in two possible orientations, nGAAnnTTCn or nTTCnnGAAn, which are determined as head-to-head or tail-to-tail orientation, respectively. With respect to complexity of the context that is difficult to analyze by the standard methods, we have used for HSE analysis an approach, which does not directly related to contextual site parameters.

Materials and methods

For analysis, we have used a sample of 42 experimentally detected heat shock transcription factor binding sites (HSE). Information about HSE is stored in TRRD database (http://www.bionet.nsc.ru/trrd/). The sites included into the sample contain extended flanking regions so that the total sequence length equals to 120 nucleotides. We have organized two subsets of the initial sample. The first subset was compiled out of the sequences aligned according the center of the binding site. The second subset was organized for analysis of conformational properties of HSE and promoter gene regions. Since it is known that heat shock genes possess by several HSE sites located in various distances from transcription start, in order to align this subset, the site sequence (-between positions –50 and +30 relatively site center) was linked with sequence carrying core promoter of this gene (between positions –30 and +30 relatively transcription start). The sample consists of 19 sequences.

The method is based on traditional approach designed for pairwise correlation analysis of amino acid sequence samples [Neher E., 1994]. This method was modified for detailed analysis of compensatory substitutions of residues in positions of sequences referring to different protein families [Afonnikov D.A. et al., 2000]. The software package for protein family analysis is available through the Internet by the address http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp/. For analyzing aligned sample set of DNA sites nucleotide sequences, we use conformational DNA properties calculated for nucleotide pairs [Ponomarenko J.V. et al., 1999]. We consider a set of N aligned sequences with the length L. In the process of analysis, each sequence is being re-coded into the sequence of nucleotides with the length L-1. Each dinucleotide is corresponded to a particular value of physico-chemical or conformational property f. Thus, we arrive at numerical matrix NxL-1. An element [l,m] of this matrix corresponds to the value of property f of a dinucleotide beginning at position m in the sequence l.

As the measure of dependency between the values of properties f in positions of the sample sequences, we take the value of correlation coefficient as follows:

$$r_{ij} = \frac{1}{N-1} \sum_{k} \frac{(x_i - \bar{x}_i)(x_j - \bar{x}_j)}{\sqrt{d_{ii} \cdot d_{jj}}}, \qquad (1)$$

where x_i is a property of a nucleotide starting at *i*– th position,

N is a number of sequences in a set,

d_{ii} is a sample variance of a property of a dinucleotide starting at *i*-th position.

Significant deviation of correlation coefficient from zero (that is, when it exceeds by modulo some limiting value $\mathbf{r_{limit.}}$) means that there exists a statistical dependency for conformational property values at a pair of positions i, j. To determine $\mathbf{r_{limit.}}$, we use the value

$$t = |r_{ij}| \cdot \left(\frac{m}{1 - r_{ij}^2} \right)^{1/2}, \qquad (2)$$

which is distributed in accordance with the Student criterion with m = N-2 degrees of freedom [Anderson T.W., 1958].

For analysis of nucleotide sequences, the problems, related to determining of statistical dependencies in a set given, do not appear due to evolutionary interrelationships between the sequences [Afonnikov D.A. et al, 2000]. This fact simplifies an application of the method.

At the subsequent stage, the task was posed to reveal the blocks with significant prevalence of correlating pairs in the matrix. The size of this block was chosen so that the number of pairs forming the square block should be essentially less than their complete number of pairs in the whole matrix. In order to determine statistical reliability of prevalence of correlating pairs in this block, we have used an approximation of binomial distribution. In this case, the probability p of event that the number of significant pairs in this block should exceed the observed value m equals to

$$p = \sum_{k=n-m}^{n} C_n^k (q)^k (1-q)^{n-k}, \qquad (3)$$

where q is a probability of significant correlation for the whole matrix,

n is the number of positions in a window.

Thus, this block was considered as significant if p < 1%, that is, this corresponds to significance level equaling to 99%.

Results and discussion

We have analyzed the correlations between the values of 38 different dinucleotide conformational properties [Ponomarenko J.V. et al, 1999]. The most interesting results for both samples were obtained under correlation analysis of persistent length values. As persistent length, a length of DNA helix is denoted, such that this helix may be accounted as a cylinder without bending and deformation.

Under analysis of both samples, it was found that for persistent length property, there exists a large bulk of pairs of positions with significant correlation coefficients. A discriminative feature is that these pairs are localized in clusters. In general, the discrepancies detected for the first subset are repeated at the corresponding region of the second subset, which has the less size. The result of analysis is given in Fig. 1. In this case, there was observed an interrelation between two regions located from both sides of the site (cluster 6). Moreover, within the limits of the 5'region, the relationships between neighboring positions is stronger (cluster 1) than between analogous positions in 3' region (cluster 2).

The helix in vicinities of HSF binding site has unbend strict BDNA-form, because under HSF binding, conformation of sites is characterized by small bend angle value (5.4°) in the region binding HSF. On the contrary, analogous values for HNF3 and ETS factors, which are referred to the same protein family, are larger [Littlefield O. and Nelson C.M., 1999]. This may be explained by the fact that HSF belongs to transcription factor family with "winged" DNA-binding domain of helix-turn-helix type. These data support the importance of this parameter for characterization of HSF binding sites functioning.

Following the dependency detected in analysis of relationships between promoter and site regions, we have made an analysis of the second subset. In this subset, the sequence of a site was linked to the sequence containing core promoter of this gene (Fig. 1). The most interesting are the clusters 7 and 8 reflecting the relationships between the core promoter regions and the flanking site regions. These regions are also slightly related within the limits of each region (clusters 3, 4, and 5) and are located respectively at positions in between –10 to –35 from the site center and in-between –10 to +10 in promoter region.

The results of the second subset analysis may be interpreted within the frames of competition model. HSP gene transcription regulation appears at the elongation stage, this fact explaining rapid cell response to the action of different stressing factors. Pausing PNA-polymerase II bound with TBP at HSP gene promoter positions almost near-by transcription start. It was demonstrated that under the heat shock action, an active HSF immediately interacts with basal transcription factor TBP [Mason P.B. Jr. and Lis J.T., 1997]. HSF competes with RNA-polymerase II for TBP binding. This, in turn, releases RNA-polymerase and activates HSP genes transcription. We suppose that for interaction of HSF and TBP factors, the 5'-flanking regions of HSE site and 3' flanking regions of TATA-box should have close spatial localization. This may be the possible reason of anticollinear or collinear disposition of these DNA regions (Fig. 2). In its turn, this may be the reason of observed interrelations between persistent length values (Fig. 1).

Conclusion

As a result of HSE sample analysis, we have revealed DNA conformational properties that may influence site functioning. These data give evidence that there exist DNA regions within site vicinities, with correlated deviation of persistance length property values. Moreover, such regularities can be found even in remote DNA regions, thus verifying possible spatial closeness of these regions, which appears due to their functional or structural peculiarities.

Acknowledgement

The work is supported by the Russian foundation for Basic Research (grants Nos 98-07-91078, 99-04-49879) and INTAS-96-1787. The authors are grateful to M.P. Ponomarenko for valuable advises, remarks, and ideas developed during fruitful discussions; to G.V. Orlova for translation of the manuscript into English.

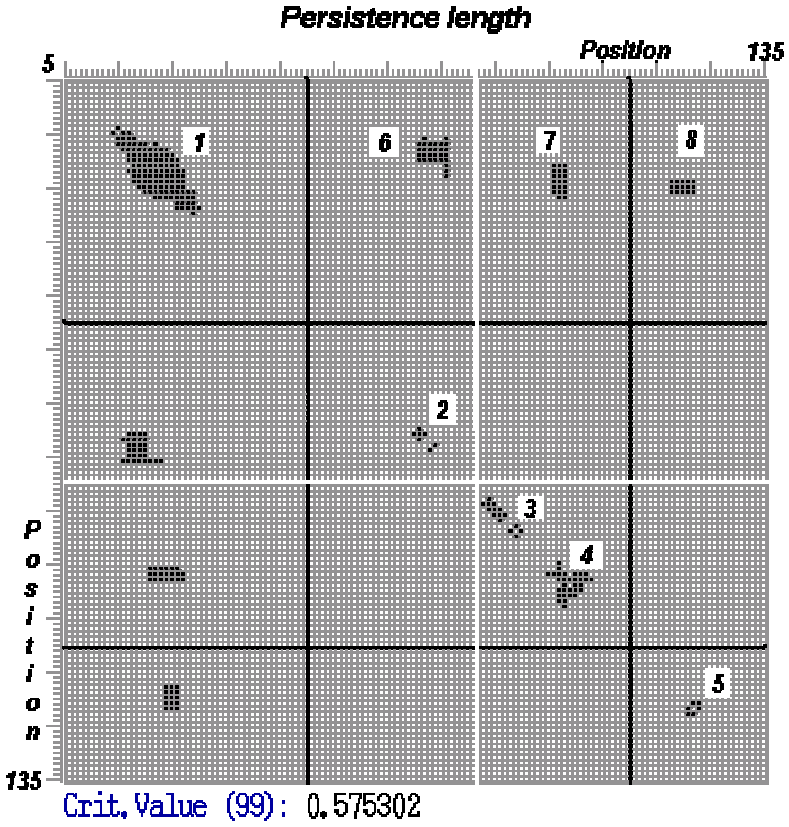


Figure 1. Positioning of significant blocks in correlation coefficients matrix calculated for persistent length values for HSE sites and promoter regions of heat shock genes between positions –30 and +30 relatively transcription start. The blocks are marked by color and numbers. By X- and Y-axes, positions of dinucleotide beginning are indicated. The value of correlation coefficient significance is estimated by the formula (2) under the confidence level of 99%. Clusterization was made for the window with the size 10x10 and confidence level of 99%, calculated by the formula (3). By white lines is marked the location where the sequences were linked. By black lines are given positions of HSE sites centers and transcription starts.

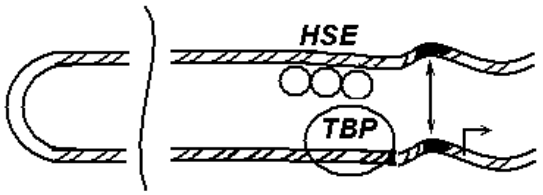


Figure 2. Hypothetical model of interactions between HSF and TBP along heat shock gene promoters. This model explains the relationships between conformational properties of remote DNA regions. In order these transcription factors could interact, DNA helix should loop in such a manner that the vicinities of binding sites are closely positioned and, possibly, anticollinear (or collinear) due to possible interactions between protein complexes. These facts may explain relatedness of conformational properties throughout the regions studied.

References

1. Anderson T.W. An introduction to multivariate statistical analysis. John Wiley & Sons Inc. 1958. NY
2. Neher E. How frequent are correlated changes in families of protein sequences? Proc. Natl. Acad. Sci USA. 1994. 91. P. 98-102.
3. Afonnikov, D.A., Oshchepkov D.Yu., Kolchanov N.A. Estimation of variances and covariances of protein physico-chemical characteristics in families of homologous sequences. Computational Technologies, in press (2000).
4. Ponomarenko J.V., Ponomarenko M.P., Frolov A.S., Vorobyev D.G., Overton G.C., and Kolchanov N.A., Conformational and physicochemical DNA features specific for transcription factor binding sites. Bioinformatics. 1999. 15, 7/8. P. 654-668.
5. Mason P.B. Jr, Lis J.T. Cooperative and competitive protein interactions at the hsp70 promoter. J. Biol. Chem. 1997. 272(52) P.33227-33233
6. Littlefield O., Nelson H.C. A new use for the 'wing' of the 'winged' helix-turn-helix motif in the HSF-DNA cocrystal. Nat. Struct. Biol. 1999. 6(5) P.464-470

SINGLE NUCLEOTIDE POLYMORPHISM IN THE REGION OF 288-296 BP OF INTRON 2 OF THE K-RAS GENE, RELATED TO LUNG TUMOR SUSCEPTIBILITY, CAUSES ALTERATION IN THE SET OF PROTEINS BINDING TO THIS REGION

Levashova Z.B., Kaledin V.I., Ponomarenko M.P., Kobzev V.F., Vasiliev G.V., Ponomarenko J.V., Podkolodnaya O.A., *Merkulova T.I., Kolchanov N.A.

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

Keywords: site recognition method, cluster analysis, GATA transcription factor, protooncogen K-*ras*, single nucleotide polymorphism, tumor

Resume

Motivation:

Single nucleotide polymorphism within the region 288-296 bp of the 2 intron of the mouse K-*ras* gene is related to different susceptibility of inbred mouse strains to spontaneous and chemically induced lung tumor [Chen B. et al., 1994].

Result:

By applying mobility shift assay, we have demonstrated that the oligonucleotide corresponding to “sensible” variant of single nucleotide polymorphism CA, reveals an additional complex with the proteins of lung cells nuclear extract in comparison to oligonucleotides reproducing “stable“ GC and “intermediate“ CC variants. By the method recognizing transcription factor binding sites and by competition with oligonucleotides corresponding to the known regulatory elements the protein binding to sensible allele was identified as a member of the GATA family.

Availability:

MATRIX database, URL= http://wwwmgs.bionet.nsc.ru/mgs/systems/consfreq/.

Introduction

Recognition of transcription factor binding sites is one of important components of computer-assisted analysis of different gene nucleotide sequences. Besides, it is very productive approach for solving many problems in experimental molecular biology. One of perspective applications of the methods designed for recognition of these sites is revealing of functional significance of single nucleotide polymorphisms (SNP) in non-coding gene regions. Notably, SNP studies related to inherited susceptibility to various pathologies is of especial significance.

The K-*ras* gene is used as a genetic marker of susceptibility in different mouse strains to spontaneous and chemically induced carcinogenesis in lung [1]. Three alleles of this gene are known. They are denoted as sensible (K^s), intermediate (K^i), and resistant (K^r) alleles and are related to different expression patterns of this gene. All K^r allele carriers are characterized by tandem repeat of 37 bp in length in the second intron (282-355 bp). K^s and K^i carriers have only a single copy of this repeat. In addition, there are two single nucleotide substitutions inside the repeated unit, which correlate to lung tumor susceptibility. In particular, the sensible allele has the C nucleotide at 288 bp position and A nucleotide at 296 bp position, whereas the intermediate allele is characterized by substitution A→C, and resistant allele – by substitutions C→G and A→C, correspondingly [2]. We have supposed that nucleotide substitutions may be located within the region binding to some regulatory protein, thus leading to decrease in the K-*ras* gene expression. To testify this hypothesis, we have analyzed the binding of lung cells nuclear extract proteins to synthesized double-strand oligonucleotides corresponding to all three allelic states of single point polymorphism (CA, CC, and GC) in the 2 intron of the K-*ras* gene. Besides, computer analysis of these oligonucleotides by the methods designed for recognition of transcription factor binding sites was made.

Materials and Methods

We have studied 3 oligoDNA, of 30 bp in length, namely, "K-ras-CA"=gtgcaagaaa**C**tccactt**A**tcatgagagct, "K-ras-CC"=gtgcaagaaa**C**tccactt**C**tcatgagagct, and "K-ras-GC"=gtgcaagaaa**G**tccactt**C**tcatgagagct, corresponding to three allelic states within the region in-between 278-307 bp of the 2 intron of the mouse K-*ras* gene. For both

DNA strands, we have determined the maximum of the function for recognition of transcription factor binding sites by constructing the oligonucleotide frequency matrix [3]. In total, we have analyzed 467 matrices, which were constructed on the base of 1458 known sites referring to 41 types (Table 1, columns I-IV). Analysis of the results obtained was made by the standard statistical software package STATISTIKA.

Results and Discussion

Binding of nuclear proteins to double-strand oligonucleotides corresponding to three allelic states and located within the region 278-307 bp of the second intron of the mouse K-*ras* gene was studied by mobility shift assay. Lung cell nuclear extract proteins were shown to form 3 complexes with the oligonucleotide corresponding to "sensible" variant of two-point polymorphism CA (Fig.1). In case of oligonucleotides producing "stable" GC and "intermediate" CC variants, one of the main bands of retardation has disappeared from the gel (complex 3). Since the variant CA is related to increased expression of the K-*ras* gene, it may be suggested that the protein producing typical for this variant retardation band is capable to activate transcription of the K-*ras* gene. For identification of this unknown protein, we have used the methods for transcription factor binding sites recognition, the results of which are presented in Table 1 (columns V-X). In order to detect the protein-candidate, which has its binding site in the K-ras-CA, we have introduced the patterns X(+) and X(-), of site recognition function (Table 1: lines I and II). In addition, the pattern "None" was introduced, it denotes the function recognizing protein site which does not bind to K-*ras* (Table 1, line III). Then the similarity of transcription factor TF (Table 1, lines 1-41) with the X-pattern is expressed as follows:

$$\Delta_{TF,P} = D_{None,X} - D_{TF,X} - t_{\alpha=0,05;\nu=40} \times s.d.(D_{TF,X}) \quad (1)$$

.where $D_{Q,P}$, Euclidean distance between the "Q" pattern and this pattern "P" at the columns V-X of Table 1 (); t, Student's coefficient; s.d., standard deviation (Table 1, columns XI-XIII).

Positive $\Delta_{TF,P}$ values denote significant (α<0,05) similarity of the factor TF to the pattern K-ras binding to "P" rather than to the pattern "None". The results given by the Eq. 1 can be seen in the columns XI-XII of the Table 1. Really, the only positive value of $\Delta_{GATA,X(-)}$=0,33 points to the GATA factor as the most probable candidate, which have the binding site in (-)-strand of K-ras-CA variant. To the other possible candidate, c-Myb, binding to (+)-strand of the K-ras-CA variant, indicates the estimate $\Delta_{c\text{-}Myb,X(+)}$=-0,06, which is the highest out of the negative estimates. By analogy, the similarity of TF/None-patterns equals to

$$\Delta_{TF,None} = - D_{TF,None} + t_{\alpha=0,05;\nu=40} \times s.d.(D_{TF,None}). \quad (2)$$

Positive $\Delta_{TF,None}$ values (Eq. 2) indicate to similarity with the pattern "None" (Table 1, column XIII). Out of 41 proteins studied, 13 proteins, i.e., AP-1, ATF, c-Jun, COUP, CP-1, CRE-BP1, CREB, E2, NF-IL6, OCT, RAR, SRF, and USF have $\Delta_{TF,None}$>0, whereas their binding to K-ras looks as the lest probable. For independent control of predictions obtained by Eq. (1 and 2), we have analyzed the values from the Table 1 (columns V-X) by means of the standard statistical package STATISTICA.

An example of the resulted data produced by STATISTIKA package is given in Fig. 2: the GATA factor was found as the most similar to the pattern X(-), whereas the c-Myb factor was mostly resembling the X(+) factor. Besides, 12 out of 13 non-binding K-ras factors (except CP-1) were found to be localized in the same cluster as "None" pattern. In the Table 2 illustrating combinations of 7 cluster-analysis methods and 5 similarity scales, one can see the factors mostly resembling the patterns X(+) and X(-): in all 35 cases considered, GATA factor was the nearest to X(-), whereas c-Myb factor was the nearest to X(+) only in 32 cases.

Thus, the results of the software package STATISTIKA implementation (Table 2 and Fig.1) are in a good accordance with predictions given by the Equations (1) and (2): the GATA factor is the most probable candidate for binding of the K-ras-CA, and the c-Myb factor is the second reliable candidate. Binding of K-ras-CA to the factors AP-1, ATF, c-Jun, COUP, CP-1, CRE-BP1, CREB, E2, NF-IL6, OCT, RAR, SRF, and USF looks as the lest probable.

Experiments using the oligonucleotide corresponding to the known binding site of the GATA factor may serve as a confirmation of predictions made by Equations (1 and 2) and the package STATISTIKA. As can be seen from the data shown in Fig. 1, addition of 50-fold excess of the oligonucleotide carrying the GATA binding site, but not the control oligonucleotide, leads to complete disappearance of the complex 3.

Table 1. Search for the TF-candidate which binding site on K-ras DNA was damaged by SNP's.

TF site recognition tools				TF-similarity score on K-ras mutants						Cluster analysis			Result
Factor		Data size		K-ras-CA		K-ras-CC		K-ras-GC		$\Delta_{TF,X}$		$\Delta_{TF,None}$	
No	Name	N_T	N_M	(+)	(-)	(+)	(-)	(+)	(-)	X(+)	X(-)		
I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV
1	AP-1	74	16	-0,30	-0,35	-0,27	-0,32	-0,57	-0,15	-0,74	-0,74	0,29	None
2	ATF	28	10	-0,45	-0,65	-0,40	-0,15	-0,32	-0,17	-0,88	-1,00	0,17	None
3	C/EBP	126	18	0,00	-1,15	-0,18	-1,17	-0,05	-0,80	-0,91	-1,94	-0,75	
4	c-Fos	21	10	0,05	-0,37	0,25	-0,37	0,20	-0,35	-0,29	-1,02	-0,27	
5	c-Jun	32	10	-0,25	-0,15	-0,15	-0,07	-0,10	-0,15	-0,60	-0,52	0,13	None
6	c-Myb	19	10	0,75	-0,57	0,50	-0,57	0,40	-0,22	*-0,06*	-1,61	-0,98	*Probable*
7	COUP	18	10	-0,65	-0,37	-0,47	-0,35	-0,42	0,05	-1,09	-0,79	0,07	None
8	CP-1	51	10	0,05	-0,42	-0,10	-0,37	0,03	-0,37	-0,25	-0,94	0,02	None
9	CRE-BP1	26	10	-0,40	-0,15	-0,57	-0,02	-0,55	-0,07	-0,95	-0,53	0,04	None
10	CREB	46	10	-0,27	-0,57	-0,35	-0,60	-0,27	-0,47	-0,69	-1,07	0,20	None
11	E2	20	10	-0,42	-0,57	-0,32	-0,60	-0,45	-0,65	-0,89	-1,13	0,08	None
12	E2F	12	10	-0,85	-0,75	-0,45	-0,52	-0,92	-0,52	-1,46	-1,38	-0,35	
13	EN	12	10	-0,50	0,05	-0,60	-0,47	-0,70	-0,50	-1,10	-0,60	-0,06	
14	ER	31	10	-0,72	-0,45	-0,85	-0,37	-0,42	-0,37	-1,26	-1,00	-0,08	
15	Ets	16	10	0,00	-0,92	0,35	-0,90	0,30	-0,75	-0,73	-1,78	-0,75	
16	GAGA	7	10	-1,00	-0,22	-0,90	0,20	0,12	0,20	-1,61	-0,92	-0,65	
17	GAL4	16	10	0,20	-0,77	0,15	-0,80	0,32	-0,60	-0,41	-1,57	-0,59	
18	**GATA**	**81**	**16**	**-0,25**	**1,00**	**-0,62**	**0,35**	**-0,55**	**0,25**	**-1,48**	**0,10**	**-1,04**	**BEST**
19	GR	63	16	-0,75	-0,19	-0,40	-0,05	0,01	0,00	-1,14	-0,63	-0,11	
20	HNF1	41	10	-0,35	-0,47	-0,52	-0,57	-0,60	-0,80	-0,95	-1,12	-0,05	
21	HNF3	14	10	-0,40	-0,37	0,20	-0,47	-0,02	-0,55	-0,73	-0,99	-0,07	
22	HSF	7	10	-0,85	-0,40	0,35	-0,07	0,05	-0,07	-1,21	-1,01	-0,41	
23	IRF-1	6	10	-0,70	0,30	-0,42	0,55	-0,02	0,60	-1,54	-0,51	-0,91	
24	MEF-2	12	10	-0,42	-0,50	-0,80	-0,50	-0,85	-0,35	-1,14	-1,09	-0,16	
25	MyoD	17	10	0,25	-0,55	0,20	-0,47	0,01	-0,42	-0,11	-1,20	-0,30	
26	NF-1	101	18	-0,82	-0,85	-0,90	-0,85	-0,90	-0,62	-1,65	-1,65	-0,64	
27	NF-E2	12	10	-0,55	-0,22	-0,90	-0,17	-0,37	-0,17	-1,13	-0,71	-0,07	
28	NF-IL6	22	10	-0,45	-0,77	-0,42	-0,52	-0,35	-0,38	-0,91	-1,21	0,09	None
29	NF-kB	39	10	-0,80	-0,42	-0,50	-0,27	0,35	-0,22	-1,19	-1,01	-0,26	
30	OCT	100	16	-0,35	-0,05	-0,42	-0,10	-0,37	-0,07	-0,81	-0,39	0,14	None
31	PR	23	10	-0,37	-0,43	-0,40	-0,40	0,02	0,20	-0,81	-0,86	-0,05	
32	RAR	16	10	-0,37	-0,07	-0,35	-0,17	-0,03	-0,05	-0,76	-0,45	0,09	None
33	RF-X	12	10	-0,45	-0,85	-0,47	-0,57	-0,70	-0,52	-1,07	-1,37	-0,14	
34	RXR	21	10	-0,30	-0,82	-0,18	-0,75	-0,17	-0,20	-0,77	-1,31	-0,10	
35	Sp-1	197	27	-1,15	-0,30	-1,17	-0,07	-0,97	-0,41	-1,93	-1,24	-0,77	
36	SRF	29	10	-0,25	-0,05	-0,57	-0,50	-0,35	-0,25	-0,74	-0,54	0,17	None
37	T3R	22	10	-0,52	-0,09	-0,45	0,00	-0,15	0,02	-0,97	-0,45	-0,01	
38	TCF-1	6	10	-0,75	-0,11	-0,80	-0,20	0,27	-0,20	-1,25	-0,76	-0,32	
39	TTF-1	7	10	-0,47	0,05	-0,45	0,10	-0,52	-0,22	-1,01	-0,34	-0,04	
40	USF	28	10	-0,25	-0,22	-0,15	-0,15	-0,03	0,03	-0,61	-0,60	0,05	None
41	YY1	27	10	-0,85	-0,02	-0,77	-0,02	-0,22	0,06	-1,39	-0,56	-0,31	
TOTAL		1458	467	Means= - 0,33						s.d.			
Pattern		Bendability		Strong		Unknown		Unknown		0,41	0,43	0,35	
I	X(+)	(+)-chain		**1,00**	-0,33	**0,00**	-0,33	**0,00**	-0,33	0,71	-1,31	-0,81	
II	X(-)	(-)-chain		-0,33	**1,00**	-0,33	**0,00**	-0,33	**0,00**	-1,27	0,68	-0,81	
III	None	None		-0,33	-0,33	-0,33	-0,33	-0,33	-0,33	-0,69	-0,73	0,60	

Note: N_T, the set size of TF site sequences used; N_M, the set size of the oligonucleotide frequence matrixes used; (+)-chain; (-)-chain; $\Delta_{TF,X}$, Formula (1); $\Delta_{TF,None}$, Formula (2); -0,33, mean recognition score at 41 TF types and both chains of K-ras mutants, sd, standard deviation.

Table 2. Cluster-analysis for the TF vs X-pattern scores on K-ras mutants by STATISTICA-package.

Clustering	Similarity Scale				
Method	D2	ED	MD	CD	W
SL	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←HSF X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA
CL	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA
UPGA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←HNF3 X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA
WPGA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←HNF3 X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA
UPGC	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA
WPGC	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA
WM	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA	X(+)←c-Myb X(-)←GATA

Methods: SL, Single Linkage; CL, Complete linkage; UPGA, Unweighted Pair-Group Average; WPGA, Weighted Pair-Group Average; UPGC, Unweighted Pair-Group Centroid; WPGC, Weighted Pair-Group Centroid; WM, Ward's Method. **Scale:** D2, Squared Euclidean Distance; ED, Euclidean Distance; MD, Manhattan Distance; CD, Chebyshev Distance; W, Power; ←, NEAREST.

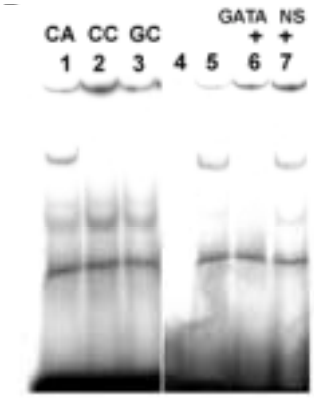


Figure 1. Binding of lung cells nuclear proteins to the oligonucleotides. 1,5,6,7 - oligonucleotide K-ras-CA corresponding to "sensible" allele of mouse K-*ras* gene; 2 – oligonucleotide K-ras-CC – to "intermediate" allele; 3 – oligonucleotide K-ras-GC – to "resistant" allele, 6 – in the presence of 50-fold excess of an oligonucleotide containing GATA binding site; 7 – in the presence of 50-fold excess of non-specific oligonucleotide, 4 – unbound probe (oligonucleotide K-ras-CA).

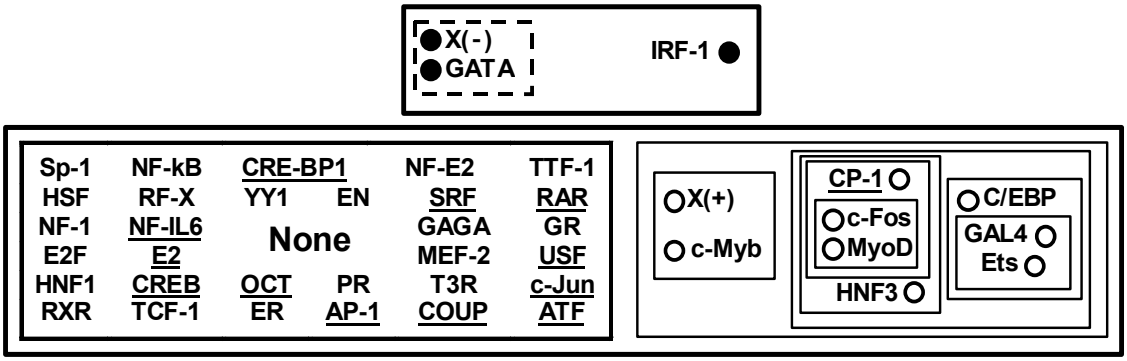


Figure 2. Example of the cluster analysis of 41 TF sites and 3 X-patterns by using the maximized TF site recognition scores upon the K-ras mutants studied by the method "Unweighted pair-group average, UPGA" and "Euclidean Distance" similarity scale.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grants Nos 00-04-49548, 98-07-90126, 98-07-91078). The authors are grateful to G. Orlova for translating the manuscript into English.

References

1. Ryan J., Barker P.E., Ruddle F.H. (1987) J. Natl. Cancer Inst. **79**, 1351-1357.
2. Chen B., Johanson L., Weist J.S., et. al. (1994) Proc. Natl. Acad. Sci. USA, **91**, 1589-1593.
3. Ponomarenko M.P., et al. (1999) *Bioinformatics*, **15**, 631-643.

REGULATORY GENOMIC SEQUENCES: CODING, ORGANIZATION, AND FUNCTION

Kolchanov N.A.

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

Keywords: genome, regulatory sequences, functional sites, coding, function. nucleosome, transcription, translation, computer analysis, recognition

Resume

Regulatory sequences (RS) control numerous functions of the genome, including all the stages of eukaryotic gene expression - transcription, splicing, translation, etc. To investigate RS, we have developed the computer system GeneExpress [1] - a superlarge integrated complex of databases and software for RS analysis and recognition, utilizing knowledge discovery and data mining (KD&DM) methods. The information on significant RS contextual, conformational, and physico-chemical characteristics discovered is stored in specialized knowledge bases with GeneExpress. This work considers certain principles of RS coding, organization, and function basing on the knowledge accumulated in GeneExpress. A more detailed information on the problems discussed is available in GeneExpress knowledge bases at http://wwwmgs.bionet.nsc.ru/mgs/systems/geneexpress/

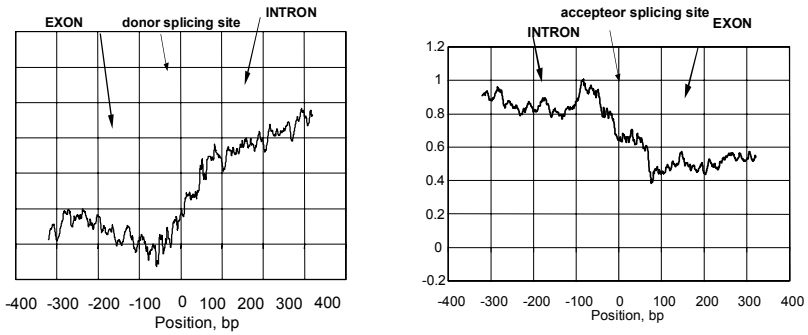


Figure 1. Profiles of the nucleosome recognition potential within the human splicing sites [6].

DNA packaging in nucleosomes and exon–intron structure of eukaryotic genes. Histone octamer is an ideally adapting molecular device capable of interacting with various DNA sequences. Analysis of nucleosomal sites allowed essential patterns of nucleosomal code to be detected [2-6]. Dinucleotides distributed along the site as to provide the double helix conformation optimal for its interaction with core histones are its elements. This nucleosomal code is degenerate: essentially different DNA sequences are capable of interacting with histone octamer. Lack of the minimal signal set in a DNA region prevents nucleosome formation.

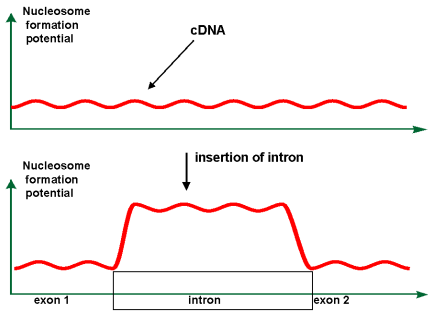


Figure 2. Insertion of intron into the gene coding region allows an efficient nucleosome formation site to be installed [8].

The developed discriminant analysis–based [6,7] method for nucleosomal site prediction allowed the patterns of nucleosome packaging in functionally differing regions of the eukaryotic genome to be investigated. A drastic increase in the nucleosome formation potential at the exon–intron boundary, its constant level within intron, and drastic decrease at the intron–exon boundary were demonstrated [5,6] (Fig. 1). A low nucleosome formation potential of exons seems to stem from their coding load, hindering location of efficient nucleosomal sites. Introns, lacking genetic code, can readily recognize the nucleosome formation signals.These results confirm the hypothesis [8] on nucleosomal chromatin organization as an

Figure 3. Nucleosome formation potential of P1 element insertion site [9].

evolutionary motivation for intron arising. It implies that whenever the primary protein structure prevented the nucleosome formation signal in the gene, an intron housing an efficient nucleosome positioning site was inserted (Fig. 2).

Insertion loci of mobile genetic element (MGE) and DNA nucleosomal organization. Typical of the genomes is uneven distribution of MGE insertion sites. The signals determining location of these sites are yet unknown. It have demonstrated that a decreased nucleosome formation potential is typical of several MGE insertion sites [9] A nucleosome formation potential profile (Fig. 3) has its minimum directly in the R1 element insertion site. We suggest that DNA nucleosome packaging might be an essential element coding for MGE insertion sites.

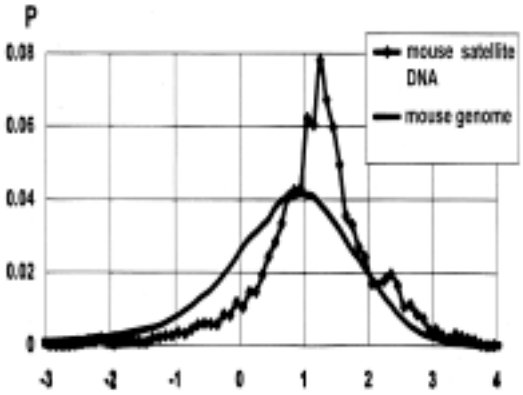


Figure 4. Nucleosome formation potential of the mouse α-satellite DNA [9].

Repeated sequences and local features of DNA nucleosomal organization. Repeated sequences and chromatin nucleosomal organization are interrelated in a peculiar way - they have either decreased or high nucleosome formation potential. Human α-satellite exemplifies those with high potential [9] (Fig. 4). Thus, a change in α-satellite copy number might modulate chromatin nucleosomal organization, thereby affecting gene expression.

Transcriptional specificity of eukaryotic promoters and DNA nucleosomal packaging. Study of vertebrate promoters has demonstrated that averaged nucleosome formation potential has a pronounced minimum at the transcription start site [7]. Further studies discovered fine nucleosomal packaging patterns of promoters with differently expressed genes: housekeeping and related genes, expressed in a wide range of tissues, exhibit an essentially lower nucleosome formation potential compared with tissue-specific genes and nucleosomal sites found experimentally [7, 10]. First of all, the promoters with housekeeping genes and genes of a wide expression range are constantly ready for transcription, as indicated by loosened (or lacking) nucleosome packaging (Table 1). This might provide for a rapid transcription initiation of such genes and its high level due to an easy access of the basal transcription machinery proteins to their promoters.

Table 1. Nucleosome recognition potential for promoters of genes differing by expression specificity [10].

Name of promoter class	Number of promoters	Average value
Housekeeping' genes	32	-1.48±0.04
Genes expressed in a wide range of tissues	30	0.66±0.04
Tissue-specific genes	141	+0.70±0.007

Distribution values are calculated for the region [-50; +1] of promoters.

As for tissue-specific genes, their promoter nucleosome packaging might be an essential element with the transcription regulation mechanism. In this case, closed chromatin is a norm (Table 1). Transcription factors, appearing in the nucleus and interacting with chromatin to loosen the nucleosome packaging, open the necessary access. Thus, tissue-specific genes have finer transcription initiation mechanisms, with changing the promoter nucleosome packaging and their repressed to activated transition as a necessary elements.

Interestingly, TATA boxes significantly differ in certain conformational properties from nucleosomal sites [11]. As is evident from Fig. 5, TATA boxes display significantly lower values of TWIST angle than random sequences; nucleosome binding sites, higher. Thus, TATA-box nucleotide context underlies certain B-DNA conformational features preventing efficient nucleosome interaction with transcription starts along with providing an easy access of TBP while initial basal transcription machinery formation. This indicates a particular importance of nucleosomal packaging in the TATA-box region for transcription initiation.

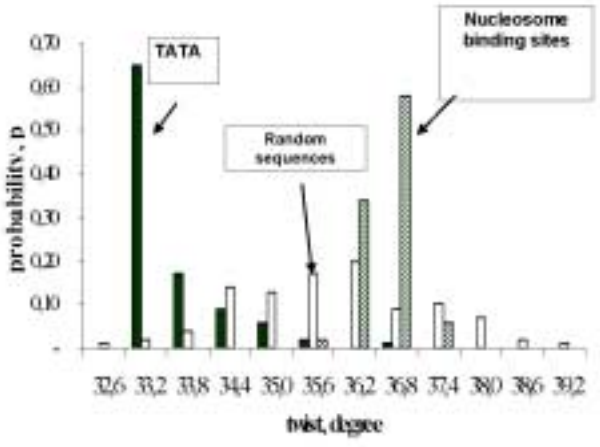


Figure 5. Distributions of the mean value of helical twist angle for the invertebrate TATA-boxes, random sequences and nucleosome binding sites [11]

Structural similarities in N- terminal domains of two TAFs, Drosophila basal transcription machinery proteins, and H3/H4 heterodimer or DNA-binding

domain of rat hepatocyte nuclear factor 3-γ and globular domain of linker histone H5 [12] suggest the interrelation of DNA nucleosome packaging and transcription initiation mechanisms. Such similarity might provide for competition between DNA-binding domains of certain transcription factors and core histones, rearranging or destroying promoter nucleosome packaging to switch them into a transcriptionally active state.

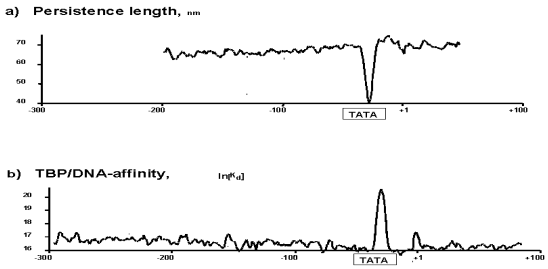


Figure 6. Examples of promoter large-scale characteristics. a - Profile of DNA bending stiffens; b - Profile of TBP affinity for promoter DNA; + 1 - transcription start position [[1,14].

Large-scale characteristics of regulatory sequences. Typically, RGS analysis covers only local site characteristics within the DNA/RNA regions of about several dozens base pairs interacting directly with regulatory proteins. We have demonstrated that large-scale characteristics formed by many hundreds base pairs are also typical of regulatory regions, in particular, promoters and eukaryotic mRNA 5'-untranslated regions (5'-UTR). Essential for these RS function is the scanning mechanism, when the regulatory protein (complex) moves along the sequence at the recognition stage.

An example is the calculated profile of TBP affinity for promoter DNA averaged over TATA+ eukaryotic promoters with a high, narrow affinity peak coinciding with TATA box, flanked by monotonically decreasing regions [1] (Fig. 6a). It is known that TBP recognizes TATA box through a one-dimensional diffusion along promoter [13]. Such affinity profile with a single pronounced maximum guarantees a high reliability and precision of TATA box recognition while assembling the basal transcription machinery. Thus, eukaryotic promoters have been the targets of negative selection directed against false TATA boxes within extended regions flanking genuine TATA boxes.

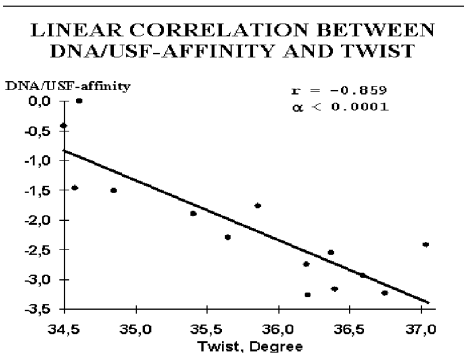


Figure 7 Correlation between the DNA/USF-affinity and the mean value of the helical twist angle of USF binding site [1].

Large-scale characteristics of regulatory regions are not only of contextual, but also of conformational nature [1,14] (Fig. 6b). For example, promoter DNA bending stifenns profile has a single pronounced minimum in the region of TATA box. It is known that TBP bends DNA while interacting with TATA box. Such bending stiffness profile provides for additional recognition accuracy, since promoter contains the single readily bending DNA region. In this process, meeting simultaneously two conditions is required for correct TATA box recognition by TBP: the site of increased DNA affinity for TBP coinciding with minimal DNA bending stiffness.

Obligatory and facultative characteristics of regulatory sequences. Specific RS interactions with regulatory proteins, determined by their contextual, conformational, and physico-chemical properties, specify RS function mechanisms and activity values. Independently of their nature, the significant RS characteristics fall into two groups [15].

Obligatory (invariant) characteristics, similar for all the RS of a given type, form the first group. These characteristics are indispensable for performing RS molecular functions and determine essentially their basal activity level. For example, dinucleotide AG is among few obligatory contextual characteristics of the splicing acceptor site. RNA molecule is cleaved at the phosphodiester bond between A and G nucleotides (explaining invariance of this dinucleotide). Specific obligatory characteristics of extended RS are exemplified by cap site, AUG codon, and local context of translation start of eukaryotic mRNA 5'-UTRs.

Facultative characteristics, individual for each particular RS in their number and location, form the second group. They provide for RS function specificities, including activity modulation of a particular RS with respect to the basal level, and are of three types-contextual, conformational, and physico-chemical. For example, transcription factor USF affinity for DNA is determined by a facultative conformational characteristic-mean value of the B-

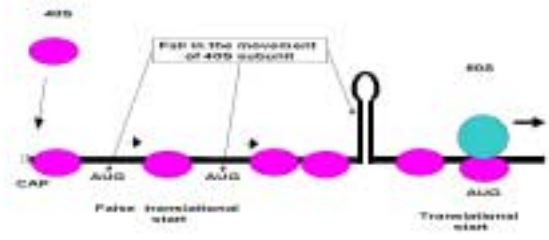


Figure 8. Moving of a ribosome along 5'-untranslated region is inhibited by false translation starts and stable hairpins.

DNA helica; twist angle, whose value depends on the USF site nucleotide sequence and encoded by a context-dependent conformational code of the B-DNA double helix [http://wwwmgs.bionet.nsc.ru/mgs/systems/activity/, 16] (Fig. 7).

It is known that mRNA translation initiation proceeds through a scanning mechanism: 40S ribosome subunit recognizes the cap site and moves along 5'-UTR to the first AUG codon displaying a necessary local context to bind to 60S subunit and start the translation. Facultative characteristics of mRNA 5'-UTR include in particular [17,18] hairpin structures, hindering 40S subunit moving along mRNA 5'-UTR and false AUG codons, halting them (Fig. 8). These and a set of other 5'-UTR facultative characteristics determine the eukaryotic mRNA translational activity [17,18].

Actually, each RS has a certain set of obligatory characteristics escorted by facultative characteristics, the number of positions occupied by the latter exceeding considerably those of obligatory characteristics. This aids both the RS evolutionary potential and their fine adjustment through fixing mutations altering facultative characteristics but leaving obligatory characteristics unchanged. This might underlie the evolutionary variation of site activity with retention of the basic molecular mechanisms of its function, determined by obligatory (invariant) site characteristics.

Transcription regulation codes. Transcription regulation codes of multicellular eukaryotes are organized to provide for a great diversity of individual gene transcription pattern. In addition to the obligatory regulatory element (RE) - core promoter, 5'-regulatory region (5'-RR) of a gene may contain dozens of various type facultative RE: transcription factor binding sites, composite elements, enhancers, silencers, etc. [http://www.bionet.nsc.ru/trrd/]. Cell nucleus of a multicellular organism contains a definite set of transcription factors depending on the cell status: cell cycle stage; cell type, tissue, or organ; developmental stage; environment; etc. Their interaction with RE and basal transcription machinery forms a specific transcription complex realizing necessary transcription level of individual gene in a particular cell (Fig. 9).

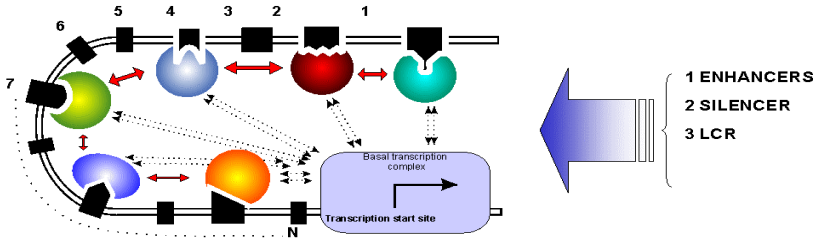


Figure 9. Gene-specific transcription complex

Each of N sites in a regulatory region exists in two states - (1) free or (2) bound to a transcription factor. In an ideal situation—absence of interfactor interactions provided - the number of states of this regulatory region equals 2(N). For example, W= 10(6) at N = 20, that is, the capacity of W code is high even in the simplest variant (0) - (1). Actually, it is by several orders of magnitude higher due to a tremendous diversity of gene-specific transcription complex versions created by protein–protein interactions between transcription factors themselves and basal transcription machinery proteins.

Typical of 5'-RR is either overlapping or inclusion of certain RE into other RE [http://www.bionet.nsc.ru/trrd/]. This pattern of genetic information coding in gene 5'-RR reflects the overlapping of contextual and conformational codes providing for alternative assemblies of specific transcription complexes.

Gene 5'-RR sizes exceed the total lengths of their coding regions by several orders of magnitude. From informational standpoint, RRs are more "loose" and less functionally loaded to the gene coding regions, manifesting in higher rates of RR evolution compared to those of coding regions. We believe this "looseness" to result from a superposition of numerous genetic messages written in codes of different natures—contextual and conformational. 5'-RR of a gene is a set of regulatory elements connected with linkers (hinges) providing individual DNA fragments with a necessary conformational freedom while forming transcription complex.

Compatibility of contextual codes is a purely informational problem. It might be solved through using different alphabets, lengths of coding words, and starting points; combining continuous genetic messages with interrupted, or positioned messages with floating; etc. Compatibility of conformational codes is an essentially more complex problem, as mechanical or physical properties of a DNA region might appear contradictory, requiring the corresponding messages in nonoverlapping regions. We believe that this is one of the reasons for a low informational load of eukaryotic gene 5'-RRs.

Positive and negative context selections in regulatory regions and functional sites. Evolutionary formation of a functional site is a multistage process involving both positive and negative selections. Positive selection fixes certain nucleotides at particular positions within the site, thereby creating the obligatory characteristics, providing for the specificity of site recognition and elementary molecular interactions determining its basal activity level, and facultative characteristics, modulating the activity with respect to the basal level [15]

The main function of negative selection is to eliminate false facultative and obligatory signals from regulatory regions and neighborhoods of sites. Our research have demonstrated that the negative selection against false signals is an important and widespread mechanism while evolutionarily optimizing the contexts of regulatory regions and functional sites controlling various genome functions and gene expression stages [15]. For example, false obligatory signals corresponding to other sites in the vicinity of a genuine site might impair its function (according to the mechanism outlined in Fig. 10).

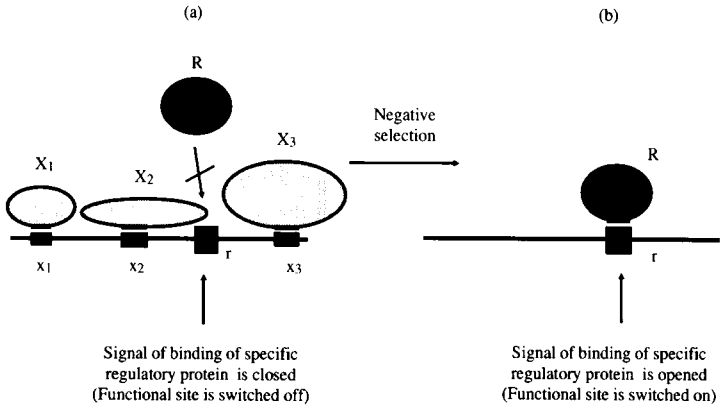


Figure 10. Negative selection against false signals in the region surrounding functional site [15]

Negative selection also plays an important role in optimizing the function of extended regulatory regions. The profile of promoter DNA affinity for TBP (Fig. 6) is a result of both the positive selection for increased affinity in a narrow region of TATA box and negative selection for the decreased affinity within TATA box–flanking regions hundreds base pairs long.

Acknowledgements

This work is partially supported by Russian State Program "Human Genome", Russian State Committee on Science and Technology, Russian Foundation for Basic Research (grants Nos 98-04-49479, 98-07-91078, 99-07-90203, 00-04-49229, 00-07-90337) and Integrational Project of SB RAS No 66.

References

1. Kolchanov N.A., Ponomarenko M.P., Frolov A.S., et al.,(1999) Integrated databases and computer systems for studying eukaryotic gene expression Bioinformatics, 15, 7/8, 669-686
2. Denisov D.A., Shpigelman E.S., Trifonov E.N. Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes. Gene 1997, 205, 145-149.
3. Ioshikhes I., Trifonov E.N., Zhang M.Q. Periodical distribution of transcription factor sites in promoter regions and connection with chromatin structure. Proc Natl Acad Sci USA, 1999, 96, 2891-2895.
4. Ioshikhes I., Bolshoy A., Derenshteyn K., Borodovsky M., Trifonov E.N. (1996) Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. J Mol Biol., 262, 129-139.
5. Levitsky V.G., Ponomarenko M.P., Ponomarenko J.V., Frolov A.S., Kolchanov N.A. (1999) Nucleosomal DNA property database, Bioinformatics, 15,7/8,582-592
6. Levitsky V.G., Kolchanov N.A. (2000) Nucleosome organization of chromatin in eukaryotic genes and structure-functional genome regions, This issue
7. Levitsky V.G., Katokhin A.V., Kolchanov N.A. (2000) Inherent modular promoter structure and its application for recognition tools development. Computation technologies (special issue), 5, 41-47.
8. Solovyev V.V., Kolchanov N.A. The eucaryotic genes exon-intron structure can be determined by the nucleosomes organisation of the chromatin and related characteristics of gene expression regulation. Dokl. Akad. Nauk SSSR, 1985, 284, 232-237(Russ).
9. Levitsky, personal communication.
10. Levitsky V.G. and Podkolodnaya O.A. (2000) Analysis of relationships between nucleosome positioning in promoter regions and gene expression pattern This issue.
11. Ponomarenko M.P., Ponomarenko J.V.et al., (1997) Computer analysis of conformational features of the eukaryotic TATA-box DNA promotors. Mol.Biol.(Mosk). V. 31 (4), P. 733-740. (Russ)
12. Nikolov D.B. and Burley S.K. (1997) RNA polymerase II transcription initiation: A structural view. Proc.Natl. Acad.Sci. USA. 94, 15-22
13. Ponomarenko M.P., Savnikova L.K., Ponomarenko J.V. et al., (1997) Modeling TATA-box sequences in eukaryotic genes. Mol.Biol.(Mosk). V. 31 P. 726-732. (Russ)
14. Babenko V.N., Kosarev P.S., Vishnevsky O.V.et al., (1999) Investigating extended regulatory regions of genomic DNA sequences, Bioinformatics, 15, 7/8, 644-653
15. Kolchanov N.A. and Lim H.A., (eds.) (1994) Computer Analysis of Genetic Macromolecules: Structure, Function and Evolution, Singapore, New Jersey, London, Hong Kong, World Scientific Pub. 556 pages
16. Kolchanov N.A., Ponomarenko M.P., Ponomarenko J.V.et al., (1998) Functional sites in pro- and eukaryotic genomes: computer models for predicting activity. Mol.Biol.(Mosk), V. 32(2), P. 255-267. (Russ)
17. Kochetov A.V., Ischenko I.V., Vorobiev D.G. et al., (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. FEBS Lett. 440, 3, 351-355.
18. Kochetov A.V., Ponomarenko M.P., Frolov A.S. et al., (1999) Prediction of eukaryotic mRNA translational properties, Bioinformatics, 15,7/8,704-712
19. Ponomarenko M.P., Ponomarenko J.V., Frolov A.S.et al., (1999), Oligonucleotide frequency matrices addressed to recognizing functional DNA sites, Bioinformatics, 15, 7/8, 631-643

SECTION 2
BIOINFORMATICS OF GENE REGULATION, GENE NETWORKS AND METHABOLIC PATHWAYS

GeneNet DATABASE: A TECHNOLOGY FOR A FORMALIZED DESCRIPTION OF GENE NETWORKS

Ananko E.A., Kolpakov F.A., *Kolchanov N.A.

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

Keywords: gene networks; database; signal transduction pathways; visualization

Resume

Motivation:

Information regarding peculiarities of gene networks organization and mechanisms of their functioning is rapidly increasing. For integration of this sort of knowledge, it is necessary to develop effective computer technology enabling to describe and visualize all integrity of elementary structures and processes in gene networks in both pro- and eukaryotes.

Results:

The computer technology for a formalized hierarchical description of a gene network has been developed. The database on gene networks (GeneNet) and the software for its automated visualization have been worked out. To provide rapid data accumulation, the Java graphic interface for the data input through the Internet is developed. Currently, GeneNet database accumulates descriptions of 18 gene networks subdivided between six sections.

Availability:

The GeneNet is available at http://wwwmgs.bionet.nsc.ru/systems/MGL/GeneNet/

Introduction

Gene network is an ensemble of coordinately expressed genes controlling vital functions [Kolpakov et al., 1998]. Regulation of the gene network operating is not restricted to the level of transcription, but it may be carried out at the levels of translation [Pyronnet et al., 1996; Buss and Stepanek, 1993], splicing [Yao et al., 1996; Pyronnet et al., 1996; Nandabalan and Roeder, 1995], posttranslational protein degradation [Hochstrasser, 1996], active membrane transport [Weissmuller and Bisch, 1993], and other processes.

The pioneer theoretical investigation of gene networks is dating back to 60-ties. These studies were devoted to consideration of general regularities of molecular-genetic system regulation in procaryotes [Ratner, 1966] and to description of gene network dynamics within the frames of the simplest logical models [Kauffman, 1969]. Later on, for the studying of gene networks dynamics, the approaches based on application of differential equations and stochastic models were suggested [Savageau, 1985; Thomas et al., 1995; McAdams and Arkin, 1997]. For integration of heterogeneous experimental information and its accumulation in databases, an effective computer technology is necessary, which permits to describe all variability of elementary structures and processes occurring in gene networks of pro- and eukaryotes within the frames of a unified approach. Moreover, it is desirable to produce an automated visualization of the gene network structure based on the formalized information stored in the database.

In the GeneNet system presented [Kolpakov et al., 1998; Kolpakov and Ananko, 1999], we have applied an original computer technology. It permits to describe any elementary structure and event occurring in pro- and eukaryotic gene networks at different hierarchical levels of organization, i.e., molecular, cellular, and referring to the whole organism levels.

Methods and algorithms

Experimental data from original papers are formalized and collected in the GeneNet database. An object-oriented approach [Booch, 1991] was employed as a basis for describing the gene network structure. For each gene network, several obligatory types of structural and functional components are marked out: 1) a gene ensemble interacting when certain biological functions are performed (the core of a gene network); 2) proteins encoded by these genes and respected for structural, transport, catalytic, regulatory, and other functions; 3) signal transduction pathways providing gene activation in response to external signals; 4) a set of positive and negative feedbacks stabilizing the parameters of the gene network (autoregulation) or providing a transition to a

new functional state; 5) nonproteinaceous substances such as signal molecules, energetic cell components, metabolites, etc. [Kolpakov et al., 1998].

All the gene network components are divided into the *Entities* (any material objects) and *Relationships* between the gene network components (Fig.1).

The entities are subdivided into 4 classes: 1) *Protein* or protein complex; 2) *Gene*; 3) *RNA*; 4) nonproteinaceous *Substance*. Instances of each class are described in a separate table in the GeneNet database The components of a gene network are scattered throughout cell compartments, cells, tissues and organs [Kolpakov et al., 1998].

Two types of relationships between the entities are considered: *Reaction*, that is, formation of a new entity or acquisition of a new property by the entity, and *Regulatory event*, that is, the effect of an entity onto a certain reaction.

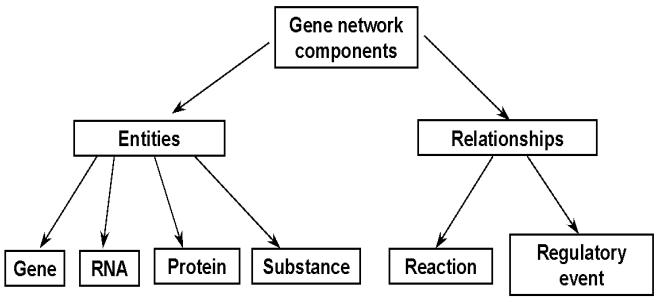


Figure 1. Classification of gene network components in the GeneNet system

The formalized description of elementary events is based on application of several informative line codes. Example is given below:

```
ID  <gene>Hs:OAS^nucleus -> <protein>Hs:OAS^cytoplasm
DT  17.5.1999; Ananko E.; created.
EF  indirect
RF  Wathelet M. et al., 1986
//
```

It means that the protein encoded by the human oligoadenylate synthase (OAS) gene is expressed in a cytoplasm (line code ID). The relationships is indirect (line code EF), because the intermediate stages such as transcription, processing, splicing, etc. are missing.

The phosphorylation initiated by the interferon receptor II (IFNR-II) of the human protein kinase Jak1 in the cell cytoplasm is described as follows:

```
ID   <protein>Hs:IFNR-II^cytoplasm ->> <protein>Hs:Jak1^cytoplasm -> <protein>Hs:Jak1-p^cytoplasm
DT  17.5.1999; Ananko E.; created.
AT  switch on
EF  indirect
RF  Silvennoinen, O. et al., 1993
//
```

Similarly, any other elementary event may be described in the terms of GeneNet, for instance, mRNA translation, enzyme catalysis, multimerization, etc.

To provide rapid data accumulation in the database, the interface for data input through the Internet was developed, which enables the user to add novel objects into the database, establish relationships between them, and to transform automatically the information into the GeneNet language (Fig. 2).

The chief merit of the technology developed is a possibility to make automated visualization of the gene network diagram. The formalized data on the gene network stored in the GeneNet database are processed by the special Java-program (GeneNet viewer) and then the data are shown up to the user as a graphical diagram (Fig. 2).

Implementation and results

All the images at the diagram are interactive, that is, if a user clicks the image, the textual description of the corresponding entry in the GeneNet database is displayed in a special text window (Fig. 2c). The text window contains a formatted text with hypertext references to other databases (EMBL, SWISS-PROT, TRRD, TRANSFAC, EPD, MEDLINE) and links with other GeneNet tables (Fig. 2).

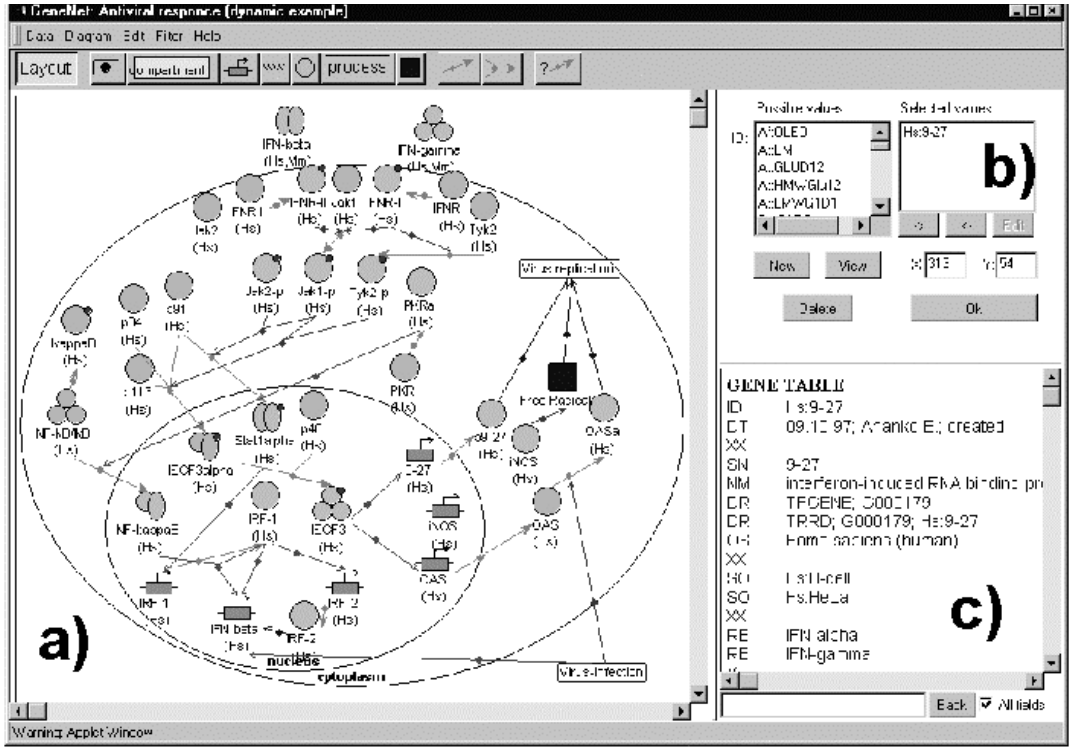


Figure 2. GeneNet graphic user interface for data input through the Internet:

a) fragment of a gene network regulating anti-viral response;
b) interactive data input window;
c) formalized description of an object (the human gene 9-27).

Informational content of GeneNet database is given in Table 1. By the 1st of April, 2000, 18 gene networks referring to 6 sections are described in the database.

Table 1. Informational content of GeneNet database (by the 1st of April, 2000).

Thematic section	Authors	Number of gene networks	Number of components		
			genes	proteins	interactions
Lipid metabolism	*Ignatieva E.V.*	2	16	22	86
Endocrine regulation	*Busygina T.V.* *Ignatieva E.V.* *Logvinenko N.S.* *Suslov V.V.*	4	31	95	187
Erythropoiesis	*Podkolodnaya O.A.*	1	32	51	87
Anti-viral response	*Ananko E.A.*	1	12	51	65
Development of seeds in plants	*Goryachkovsky T.N.* *Axsenovich A.V.*	9	45	97	390
Heat shock	*Stepanenko I.L.*	1	4	16	34
Total:		16	128	289	765

Discussion and conclusions

At present, several databases are known describing different aspects of gene network organization, e.g., CSNDB (Cell Signaling Networks Database) [Igarashi and Kaminuma T, 1997] contains an information about signal transduction mechanisms in the human cells; BRITE (Biomolecular Reaction pathways for Information Transfer and Expression) [Hopkins, 1995] accumulates the data on the cell cycle genes as well as the schemes for the pathways controlling early development in *Drosophila*; GeNet (Gene Networks database) [Serov, 1998] describes gene networks of *Drosophila*, *Nematode caenorhabditis* and *Echinus esculenta*; KEGG (Kyoto Encyclopedia of Genes and Genomes) [Kanehisa and Goto, 2000] stores the schemes of signal transduction pathways, genome maps, information about the genes; SPAD (Signaling Networks Database) (http://www.grt.kyushu-u.ac.jp/spad/) contains the structure-functional data on the mechanisms of signal transduction; EcoCyc [Karp et al., 2000] describes the metabolic pathways. However, none of these databases provides the solving of the whole complex of tasks necessary for a gene network effective studying, which demands analysis of the large bulk of heterogeneous experimental data.

The experience of this sort databases development makes clear the necessity of creating such universal computer technology that may describe any elementary structures, events, and processes significant for gene network operating. The computer technology GeneNet suggested [Kolpakov et al., 1998; Kolpakov and Ananko, 1999] is the way to the rapid accumulation of experimental data on structure-functional gene network organization, together with capacities for systematization and computer and logical analysis of this information.

Further development of the GeneNet database will proceed according to the following three directions: 1) improving of a gene network description with accounting of its hierarchical organization and spatial distribution; 2) development of approaches for mathematical modeling of gene network dynamics on the base of information stored in the GeneNet database.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (grants Nos. 98-04-49479, 98-07-91078, 99-07-90203, 00-07-90337, 00-04-49229, 00-04-49255), Russian Human Genome Program, Ministry of Science and Technology of Russian Federation, Integrated Program of the SB RAS. The authors are grateful to E.V. Ignatieva, O.A. Podkolodnaya, I.L. Stepanenko, T.N. Goryachkovsky, T.V. Busygina, A.V. Aksenovich, and V.V. Suslov for the database filling and helpful discussions; N.L.Podkolodny, and D.A. Grigorovich for SRS and software support; and G.V. Orlova for translation of the paper into English.

References

1. Buss,W.C. and Stepanek,J. (1993) Characterization of the inhibition of renal translation in the Sprague-Dawley rat following in vivo cyclosporin A. Int. J. Immunopharmacol., 15, 63-76.
2. Booch,G. (1991) Object oriented design with applications. The Benjamin/Cummings Publishing Company, Inc.
3. Hochstrasser,M. (1996) Protein degradation or regulation: Ub the judge. Cell, 84, 813-815.
4. Hopkins,T., Advani,R., Gudmunson,G. (1995) Development and implementation of an expert information system (BRITE) used in technical support of medical diagnostics customers. Clin. Chem., 41, 1333-1337.
5. Igarashi,T. and Kaminuma,T. (1997) Development of a cell signaling Network Database. Pac. Symp. Biocomput., 187-197.
6. Kanehisa,M., Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res., 28, 27-30.
7. Karp,P.D., Riley,M., Saier,M., Paulsen,I.T., Paley,S.M., Pellegrini-Toole,A. (2000) The EcoCyc and MetaCyc databases. Nucleic Acids Res., 28, 56-59.
8. Kauffman,S., (1969) Metabolic stability and Epigenesis in randomly constructed genetic net. J. Theoret. Biol., 22, 437-467.
9. Kolpakov,F.A., Ananko,E.A. (1999) Interactive data input into the GeneNet database. Bioinformatics, 15,713-714.
10. Kolpakov,F.A., Ananko,E.A., Kolesov,G.B., Kolchanov,N.A. (1998) GeneNet: a database for gene networks and its automated visualization. Bioinformatics, 14, 529-537.
11. McAdams,H. and Arkin,A., (1997) Stochastic mechanism in gene expression. Proc. Natl. Acad. USA, 94, 814-819.
12. Nandabalan,K. and Roeder,G.S. (1995) Binding of a cell-type-specific RNA splicing factor to its target regulatory sequence., Mol. Cell. Biol., 15, 1953-1960.
13. Pyronnet,S., Vagner,S., Bouisson,M., Prats,A.C., Vaysse,N. and Pradayrol,L. (1996) Relief of ornithine decarboxylase messenger RNA translational repression induced by alternative splicing of its 5' untranslated region. Cancer Res., 56, 1742-1745.
14. Ratner V.A. (1966) Molecular Genetic Regulatory Systems. Novosibirsk: Nauka. 181 p. (in Russian).
15. Savageau,M., (1985) A theory of alternative designs for biochemical control systems. Biomed. Biochim. Acta., 44, 875-880.
16. Serov,V.N., Spirov,A.V., Samsonova,M.G. (1998) Graphical interface to the genetic network database GeNet. Bioinformatics., 14., 546-547.
17. Thomas,R., Thieffry,D., Kaufman,M., (1995) Dynamical behavior of biological regulatory networks. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bul. Math. Biol. 57, 247-276.
18. Weissmuller,G. and Bisch,P.M. (1993) Autocatalytic cooperativity and self-regulation of ATPase pumps in membrane active transport. Eur. Biophys. J., 22, 63-70.
19. Yao,K.S., Godwin,A.K., Johnson,C. and O'Dwyer,P.J. (1996) Alternative splicing and differential expression of DT-diaphorase transcripts in human colon tumors and in peripheral mononuclear cells in response to mitomycin C treatment. Cancer Res., 56, 1731-1736.

PathDB: A SECOND GENERATION METABOLIC DATABASE

Mendes P., Bulmore D.L., Farmer A.D., Steadman P.A., Waugh M.E., Wlodek S.T.

National Center for Genome Resources, Santa Fé, USA
e-mail pjm@ncgr.org
*Corresponding author

Keywords: metabolic pathways; databases; pathway layout

Resume

Motivation:

Existing metabolic databases contain at least a feature that makes them unique, no single one combines *all* the desirable features. These databases are also rather limited in terms of the concept of pathways they adopt: invariably they describe pathways as well defined sequences of steps, usually the same as the textbook pathways. Additionally, none provide a user-friendly interface to allow posing complex queries and none provides any type of analysis.

Results:

Here we describe PathDB, a metabolic database developed and hosted at the National Center for Genome Resources. PathDB was designed to overcome the limitations described above. It is based on a relational schema that provides for storage of quantitative metabolic information, advanced visualization and analysis.

Availability:

PathDB is available from http://www.ncgr.org/software/pathdb

Introduction

With the advent of sequence databases in the 1980s, researchers started to construct metabolic databases. Most of these have been essentially electronic encyclopedias, with little more features than what could be found in a printed book. Here we introduce PathDB, a metabolic database that combines storage of specific and detailed information about metabolism with powerful software for data query, navigation and visualization. This combination makes PathDB more than just a simple electronic encyclopedia, enabling discovery of facts that were already known but have been hidden in the complexity of metabolic data.

Databases are computerized archives of information that usually have powerful means of indexing data. Some simply store information on disk files relying on the computer operating system for management. Relational databases organize data in tables where each row represents one single entity. Relational database management systems (DBMS) are currently the most efficient to manage large amounts of data, partly due to the powerful Structured Query Language (SQL), used to query them. Object-oriented databases organize data in a hierarchy of classes that can inherit properties from each other. Although object-oriented databases are arguably the easiest to design, their performance for large volumes of data is still far from that of relational databases.

ENZYME [1] is a database that contains the classification of enzymes by the Nomenclature Committee of the IUBMB. This has become the most used database of metabolism, perhaps because annotators of genomes like to classify enzymatic gene products according to their EC number. Apart from allowing quick searches of enzymes by EC number or vice-versa, ENZYME provides little more benefit. The pioneering work of Selkov and co-workers since the late 1980s [2] resulted in a collection of electronic information about enzymes and pathways derived from the published literature. These data are now available on the Internet in EMP [3] and MPW [4]. EMP contains quite a lot of detailed information about specific enzymes, contrasting with ENZYME that only contains classes of enzymes. EMP is a relational database that can be queried using SQL but it lacks a user-friendly mechanism for making complex queries and has limited visualization capabilities. KEGG [5] is a database of metabolic pathways that contains nice diagrams of pathways. These are static images that are updated when new steps are added to the pathway. Unlike EMP, KEGG has very little detail about the enzymes, basically not much more than the generic information that is contained in ENZYME. EcoCyc [6] is a database of *E. coli* genes and metabolism currently under control of Pangea Systems Inc. (though still free for academics). This database is based on a frame knowledge representation system, similar to an object-oriented database. Unlike the other metabolic databases, EcoCyc is specific for a single species and has benefited from curation by a domain expert. EcoCyc is original in that its graphical user-interface (GUI) includes automatic layout of metabolic pathways [7]. This is an advanced feature that we think is required if metabolic databases

are to be more than simple electronic books. However the benefits of using such technology in EcoCyc are unclear, at least in the academic version, since pathways in this system are static objects and the same diagram is always drawn for any specific one. For the user there is no difference between this or the case where the pathway diagrams are stored as static images, like KEGG. Ochs and co-workers have described [8, 9] their development of a metabolic database based on the relational model. Much in the same line that we take here, they argued that computerized metabolic maps should be useful to uncover known, but not noticed, relations between the data. However, these features are not present in the database that they have made available on the Internet, Pathways+. Finally, a collaboration between the European Bioinformatics Institute and the University of Köln has made public the enzyme data collected since 1987 by Schomburg, which is essentially an electronic form of the "Enzyme Handbook" [10]. It contains an extensive amount of data but does not have any means for making complex queries or any type of graphical visualization.

Here we describe PathDB, a metabolic database developed and hosted at the National Center for Genome Resources. PathDB has been designed to store a wide range of data in very great detail, including kinetic information; locations ranging from sub-cellular to whole organism level; taxonomic information; and thermodynamic properties of reactions. PathDB is a relational database with a non-redundant, hierarchical design. A Query Tool allows construction of complex queries and a Pathway Viewer generates pathway diagrams automatically.

Design and implementation

At the lowest level PathDB uses a relational database management system (currently Sybase v. 11.9.2). This allows the database to grow without concerns for performance issues and, more importantly, to allow powerful queries to discover relations between the data. On top of this relational database we have constructed an intermediary layer that masks it to an object-oriented view such that the user interface software sees the database as if it was object-oriented. This provides flexibility and insulates most of the software from the particularities of the DBMS used, thus minimizing the changes required if the DBMS was to change. The user-interface for PathDB consists of a suite of programs written in the Java language (v. 1.1.7 or above).

The Query Tool (QT) is the front-end for searching the database. It connects directly to a server at NCGR which provides the interface to the Sybase DBMS. The QT allows simple and wild-card queries for each of the basic data types. Results are returned as a list of objects which can then be "transformed" into other types of data. These transformations consist of retrieving objects of different nature but which are related to the original ones. For example, one can query for all compounds whose name starts with the string "Glucos%" (the '%' character is a wild-card meaning "any other characters after this position"). The result would be a list of compounds including Glucose and Glucose 6-phosphate. With that list the user has a choice of several transformations, for example "Reactions in which these compounds take part" or "Pathways in which these compounds participate", among others. Transformations allow one to navigate the complex web of relationships between metabolic entities. The QT also allows one to combine several query results using the set operations union, intersection and difference. One can thus ask questions as complex as "from what species is there evidence about enzymes that catalyze reactions involving medicarpin?".

Pathway Viewer (PV) is a component for graphical visualization of pathways. We have constructed it such that it displays pathways in a manner very similar to what biochemists use in publications. Steps are represented by arrows connecting compounds; some of the co-substrates or co-products can be classified as "secondary" and represented in a smaller font on the side of the step arrow. The "primary" compounds are displayed in a single location in the graph with all steps that produce and consume them connecting to that location. "Secondary" compounds are represented several times, as many as steps in which they take part. PV is capable of laying out pathway diagrams automatically. This has been achieved by adding, where needed, extra nodes representing the step. This transformation makes a pathway become a graph (in fact a special type called a Petri net). The user has the possibility of using any of these algorithms to draw the pathway, allowing different views of this pathway. Some of these may highlight a property of the pathway that would be difficult to discover by inspection of other diagrams. Although PV normally produces pathways arguably as good as many in publications, the user can rearrange the diagram by dragging the compounds on the screen. The program takes care of "carrying" the steps connected to the compound being moved with it so that the pathway diagram is never destroyed.

Because some users may not want to download and install a program just to query PathDB a few occasional times we have also developed a simple WWW interface. This is based on a PERL-CGI program that queries the DBMS directly and can be accessed from a Web browser. This also provides a means for adding links in web pages to specific records in PathDB. This interface, though, does not provide the advanced query capabilities that the QT does, but it is a fast way for answering simple queries (like those that can be made to the other databases mentioned earlier).

Discussion

PathDB is the result of our intention to make a database that is more than a simple electronic reference of metabolism. To carry out useful data mining of metabolic information, we believe that a metabolic database must be based on a rich data model allowing for storage of quantitative data, such as kinetics and thermodynamics of reactions, including error estimates if available. Furthermore, the objects represented in the database must be specific single biological entities. Another requirement for accurate representation of metabolism is to include transport steps and spontaneous reactions. We have also provided a data model in which every enzyme, transport protein or pathway can be labeled to be located in a specific sub-cellular and/or organ location. Enzymes and transport proteins are further subdivided into subunits which may contain a link to a gene in a sequence database and all of the latter are linked to their own taxonomic species. Naturally, PathDB links data to appropriate bibliographic references.

Given that the definition of pathways is still an open question which can easily generate disagreement among biochemists (where does a pathway begin and end?), we have decided to adopt the broadest concept for PathDB. A pathway is a set of metabolic steps (reaction or transport) connected by common intermediates, without any other requirements. This means that a single step can be seen as a pathway, as well as the whole of the database. More importantly this allows the "textbook pathways" to be represented in the database along with others that may have never been seen before. We are currently developing the software to allow users to search for routes between any two compounds in the database, to find the vicinity of compounds or steps and to visualize them. This is the main reason why we developed Pathway Viewer with automatic layout since the number of pathways contained in even a small subset of the database is enormous (all possible sets of connected reactions).

A metabolic database is an extremely useful tool to modelers of metabolism. The slowest step of constructing mathematical models of metabolism is chasing the parameter values in the literature. We plan to make PathDB export data in the Gepasi [11] file format.

A critical issue that we have not yet addressed properly in PathDB is the amount of data stored. For this database to become really successful we think that it should contain a very large proportion of the information existent in the literature. Collecting all those data into electronic format is a costly exercise that requires expert supervision. Currently NCGR is seeking funding for this large scale endeavor. Until then PathDB is focused on plant metabolism, an area that is until now poorly addressed in electronic format. However we stress that the database system are not specific to plants.

References

1. Bairoch, A. (1993) Nucleic Acids Research 21, 3155-3156.
2. Selkov, E.E., Goryanin, I.I., Kaimatchnikov, N.P., Shevelev, E.L. and Yunus, I.A. (1989) Studia Biophysica 129, 155-164.
3. Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., Jr. and Yunus, I. (1996) Nucleic Acids Res 24, 26-8.
4. Selkov, E., Jr., Grechkin, Y., Mikhailova, N. and Selkov, E. (1998) Nucleic Acids Res 26, 43-5.
5. Goto, S., Nishioka, T. and Kanehisa, M. (1998) Bioinformatics 14, 591-9.
6. Karp, P.D., Riley, M., Paley, S.M. and Pelligrinitoole, A. (1996) Nucleic Acids Research 24, 32-39.
7. Karp, P.D. and Paley, S. (1994) in Proceedings of the Third International Conference on Bioinformatics and Genome Research (Lim, H., Cantor, C. and Bobbins, R., eds.).
8. Ochs, R.S., Qureschi, A., Sycz, A. and Vorbach, J. (1996) J. Chem. Inf. Comput. Sci. 36, 594-601.
9. Ochs, R.S. and Conrow, K. (1991) J Chem Inf Comput Sci 31, 132-7.
10. Schomburg, D., Salzmann, D. and Stephan, D. (1990-1995) Enzyme Handbook, Classes 1-6, Springer.
11. Mendes, P. (1993) Computer Applications in the Biosciences 9, 563-571.

GeneNet-BASED MODEL OF TWO-STAGE ALDOSTERONE EFFECT ON PRINCIPAL CELLS OF CORTICAL COLLECTING DUCTS

**** Logvinenko N.S., Ignatieva E.V., Ivanova L.N.***

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

Keywords: aldosterone, Na+,K+-ATPase, kidney, computer simulation, GeneNet, molecular genetic model

Resume

Motivation:

GeneNet-based accumulation and visualization of data and their application for describing gene networks of molecular genetic mechanisms underlying regulation of physiological functions is an important background for further theoretical analysis of these data. Study of the gene networks mediating the effect of hormones on target cells is of special interest.

Results:

We have systematized the available experimental data on immediate (nongenomic) and slow (genomic) patterns of aldosterone effect on the target cells and we are suggesting a model of two-stage aldosterone effect on the Na^+,K^+-ATPase function in the principal cells of rat kidney cortical collecting ducts (CCD). A GeneNet-based formalization and representation of these data as a model allowed us to reconstruct the gene network comprising both the classic (genomic) mechanism of aldosterone action in cortical collecting ducts and immediate (nongenomic) realization of aldosterone function.

Availability:

The system GeneNet is Internet-accessible at http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/. The diagram of the model described is titled *Principle Cell of CCD.*

Introduction

The computer technology GeneNet, developed at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences, accumulate and visualize the data on molecular genetic mechanisms underlying the regulation of organism physiological functions [Kolpakov,F.A. et al., 1998]. The format of GeneNet provides for formalized accumulation of the information on major components of gene networks—objects (genes, proteins, low-molecular-weight substances, inducers, etc.); their interactions and interrelations; and organs, cells, and cell compartments wherein these objects are located and their interactions are realized. Java viewer allows the entire GeneNet-filed graphical and textual information on the objects within a diagram and their interactions to be accessed. Current release of GeneNet contains a number of gene networks [Kolchanov,N.A. et al., 2000]. Analysis of the GeneNet information and the available published data allow several major types of gene networks to be separated: (a) gene networks providing for cyclic processes, *e.g.* cell cycle, heart muscle constriction cycle, etc.; (b) providing for cell growth and differentiation, tissue and organ morphogeneses, organism growth and development (erythropoiesis, seed germination, etc.); (c) providing for homeostasis of organism biochemical and physiological parameters; and (d) providing for organism reactions to changes of the environment (heat shock, antiviral response, etc.).

The goal of this work was to develop a GeneNet-based model of gene network providing for the effect of hormone in its target cells. The molecular mechanism of aldosterone effect on Na^+,K^+-ATPase of kidney cortical collecting duct cells was used as the object.

Brief characteristics of biological objects

Na^+,K^+-ATPase, or the so-called sodium pump, is present in all the eukaryotic cells. It is a membrane protein consisted of α- and β-subunits, assembled at a 1 : 1 ratio to form a stable heterodimer. This enzyme is capable of drawing three Na^+ ions out and bringing two K^+ ions into the cell against their gradients per each ATP molecule used [Skou J.Ch. and Esmann M., 1992]. It is involved in forming transmembrane potential; controlling cell volume; and regulating Na^+-dependent transport of protons and other ions, sugars, and amino acids. Hormones and neurotransmitters regulate the activities of Na^+,K^+-ATPase in different tissues [Ewart H.S. and Klip A., 1995].

Principal cells of kidney cortical collecting ducts (CCD) of higher animals are the main targets of aldosterone, thus providing for a hormone-dependent regulation of sodium reabsorption.

Aldosterone is the key steroid hormone regulating the activity of Na^+,K^+-ATPase in the principal cells of cortical collecting ducts.

Reconstruction of gene network

The data on Na^+,K^+-ATPase activity in principal CCD cells are dispersed in numerous publications, each describing results of studying certain aspects or stages of this regulation. To integrate the experimental information on mechanisms of aldosterone effect on CCD cells, we have reconstructed the gene network using the system GeneNet. The information was inputted into GeneNet via the Internet using the interactive data input system [Kolpakov F.A. and Ananko E.A., 1999]. Each component of the gene network has its own image reflecting its specific features (Fig. 1a). For example, a monomeric protein is represented by a circle; homodimeric protein, by twin oval, etc. The overall diagram was constructed by uniting the available data on elementary molecular events and processes from the relevant papers. Graphic representations of elementary events and processes are exemplified in Fig 1 (b–f).

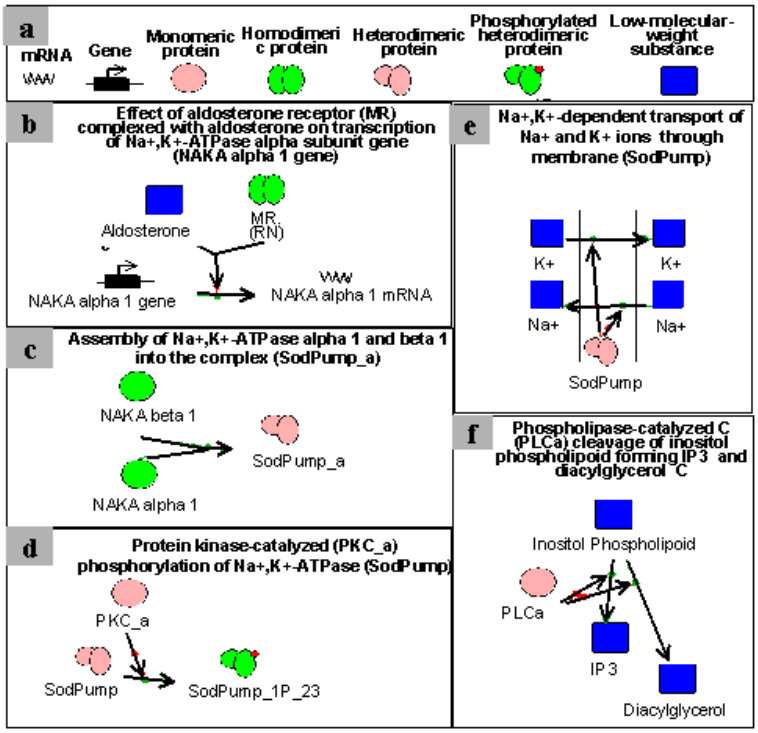


Figure 1. Representation of (a) objects and (b-f) elementary processes on GeneNet diagram.

Gene network of two-stage aldosterone effect on cortical collecting duct cells

The model with the GeneNet database comprises two mechanisms of aldosterone effect on Na^+,K^+-ATPase on principal CCD cells of animal kidneys (Fig. 2). First is a classic genomic mechanism, the so-called long pattern (*Long* in Fig. 2). It requires hours and days for manifesting the aldosterone effect on the sodium pump [Verrey,F. et al., 1996]. Intracellular transduction of the hormone signal commences from aldosterone–mineralocorticoid receptor (MR) interaction [Funder,J.W., 1992; Marver,D., 1992]. Once aldosterone binds to receptor, the complex formed is activated accompanied by dissociation of heat shock proteins (not shown in diagram). The activated hormone–receptor complex is translocated into the nucleus, interacts with binding sites of mineralocorticoid receptors located in the regulatory regions of the genes coding for α1- and β1-subunits of Na^+,K^+-ATPase (NAKA alpha 1 and NAKA beta 1 genes), and increases the syntheses of corresponding mRNAs (NAKA alpha 1 and NAKA beta 1 mRNAs) [Verrey et al., 1989; Logvinenko et al., 1991a; Logvinenko et al., 1991b; Farman et al., 1992]. This increases the number of new pumps localized into the basolateral membrane of principal cells (SodPump) and elevates consequently the sodium reabsorption and potassium secretion. Thus, this part of the model describes production of new additional Na^+,K^+-ATPase molecules and represents the traditional area of aldosterone action.

The nongenomic pattern of aldosterone action (*Fast* in Fig. 2) is represented by immediate aldosterone effect on Na^+,K^+-ATPase and mediated by secondary intermediates [Wehling,M., 1995]. This action takes minutes and seconds for its manifestation. Aldosterone binds to the membrane receptor (MRm), thereby stimulating phospholipase C activity (PLCi, inactive form; PLC, active form). Phospholipase C hydrolyzes membrane phosphoinositides (Inositol Phospholipid) to form two secondary mediators—inositol triphosphate (IP3), increasing intracellular Ca^{2+} content, and diacylglycerol [Schneider.M. et al., 1997]. Diacylglycerol and Ca^{2+} activate Ca^{2+}-dependent protein kinase C (PKC_i, inactive form; PKC_a, active form) and switch on the cascade of phosphorylation of pump preexisting molecules [Bertorello,A.M. et al., 1991; Beguin P. et al., 1994]. Protein kinase C activation (PKC_a) leads to phosphorylation in serine at position 23 of the *N*-terminal part of Na^+,K^+-ATPase α1-subunit and to a inhibition of the pump hydrolytic activity under *in vitro* conditions (SodPump_1P_23) [Feschenko M.S. and Sweadner K.J., 1995, Logvinenko,N.S. et al., 1996]. This pattern of

Na^+,K^+-ATPase regulation may be modulated by a preliminary phosphorylation of serine at position 943 in α1-subunit (the phosphorylated protein is designated as SodPump_1P_943) by cAMP-dependent protein kinase A [Belusa R. et al., 1997]. The preliminary phosphorylation of serine at position 943 by protein kinase A significantly attenuates the PKC phosphorylation effect and results in appearance of the additional phosphorylation sites located on the big intracellular loop of the α1-subunit (this form of Na^+,K^+-ATPase, phosphorylated at several positions, is designated SodPump_>2P). Note that the inverse phosphorylation order fails to exhibit such attenuation effect. The functional effect of kinase-dependent Na^+,K^+-ATPase phosphorylation depends linearly on the intracellular concentration of Ca^{2+} ions [Cheng,S.X.J. et al., 1999] (not shown on diagram). PKA and PKC stimulations at low Ca^{2+} concentration (125 nM) decrease the sodium pump activity, whereas at high concentration (450 nM), phosphorylation results in the increase of the activity .

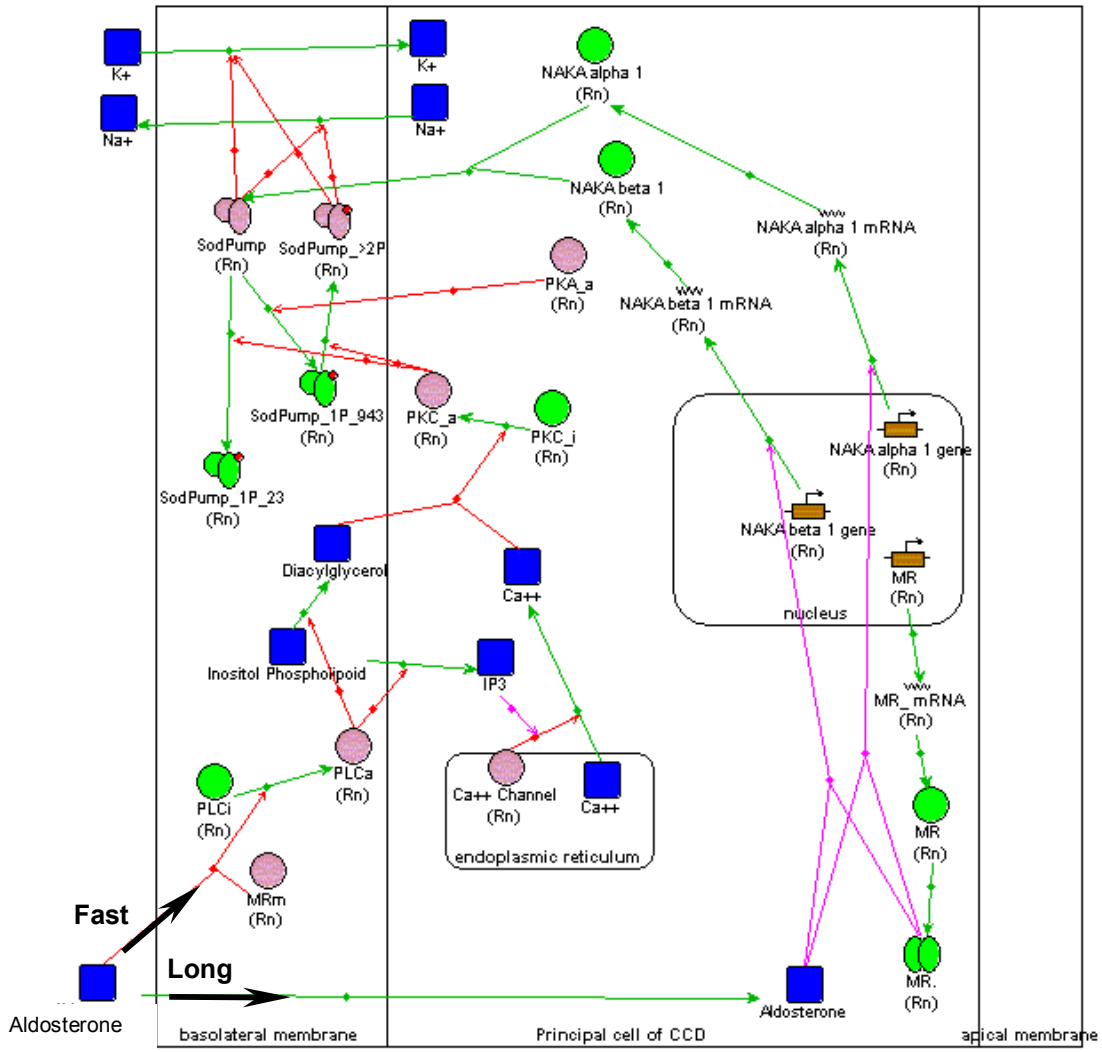


Figure 2. Graphical representation of information from the section "Principal cell of CCD" of the GeneNet database.

Conclusion

The new gene network with the GeneNet database characterized describes the mechanisms underlying realization of the hormone effect on its target cell. Descriptions of both the classic genomic and nongenomic patterns of aldosterone effect are the specific feature of the developed model of Na^+,K^+-ATPase activity regulation in principal CCD cells. Of interest is the fact that cell membrane–located proteins—aldosterone membrane receptor and phospholipase C—play the key role in realization of the immediate, nongenomic regulation pattern. According to the model, aldosterone can act through both routes of intracellular hormonal signal transduction simultaneously; interaction of these two mechanisms might provide for a wide range of cellular reactions. Insight into these processes is of considerable value from physiological standpoint, as it clarifies a variety of controversial data on intracellular aldosterone effects. Computer-based representation of the model makes it open to further development on acquisition of new experimental data. In future, we believe extremely purposeful to systematize the data on intercrossing of aldosterone effect realization routes with those

of other hormones (in particular, vasopressin) of CCD cells in the framework of this model. In addition to the advantages noted, this GeneNet-based computer model, to the knowledge of authors, is the first computer image of the molecular mechanisms underlying aldosterone effect transduction in the target cell.

Acknowledgments

The work was supported by the Russian Foundation for Basic Research (98-04-49479, 99-04-49907, and 00-15-97974). The authors are grateful to G.B. Chirikova for translation of the paper into English.

References

1. Beguin P., Beggah A.T., Chibalin A.V., Burgener-Kairuz P., Jaisser F., Mathews P.M., Rossier B.C., Cotecchia S., Geering K.*J. (*1994) Phosphorylation of the Na,K-ATPase alpha-subunit by protein kinase A and C in vitro and in intact cells. Identification of a novel motif for PKC-mediated phosphorylation. *Biol Chem* 269, 24437-24445.
2. Belusa R., Wang Zh-M., Matsubara T., Sahlgren, B., Dulubova I., Nairn A.C., Ruoslati E., Greengard P., and Aperia A. (1997) Mutation of the protein kinase C phosphorylation site on rat α1 Na+,K+-ATPase alters regulation of intracellular Na+ and pH and influences cell shape and adhesiveness. J. Biol. Chem., 272, 20179–20184.
3. Bertorello A.M., Aperia A., Walaas S.I., Nairn A.C., and Greengard P. (1991) Phosphorylation of the catalytic subunit of Na+,K+-ATPase inhibits of the activity of the enzyme. Proc. Natl. Acad. Sci. USA, 88,11359–11362.
4. Cheng, S.X.J., Aizman, O., Nairn, A., Greengard, P., and Aperia, A. (1999) [Ca2+]i determines the effects of PKA and PKC on the activity of rat renal Na+,K+-ATPase. J. Physiol., 518, 37–46.
5. Ewart H.S. and Klip A. (1995) Hormonal regulation of the Na+,K+-ATPase: mechanisms underlying rapid and sustained changes in pump activity. Am. J. Physiol., 269, C295–C311.
6. Farman N., Courty N., Logvinenko N., Blot-Chabaut M., Bourbouze R., and Bonvalet J.P. (1992) Adrenalectomy reduces α1 and not β1 Na+,K+-ATPase mRNA expression in rat distal nephron. Am. J. Physiol., 263, C810–C817.
7. Feschenko M.S., Sweadner K.J. (1995) Structural basis for species-specific differences in the phosphorylation of Na,K-ATPase by protein kinase C. *J Biol Chem* , 270, 14072-14077.
8. Funder J.W. (1992) Regulation of transepithelial Na transport by steroid and peptide hormones. Steroid receptors. Semin. Nephrol., 12, 6–11.
9. Kolchanov N.A., Ananko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., and Stepanenko I.L. (2000) Gene networks. Mol. Biol. (Mosk.), (in press).
10. Kolpakov F.A., Ananko E.A., Kolesov G.B., and Kolchanov N.A. (1998) GeneNet: a gene network database and its automated visualization. Bioinformatics, 14, 529–537.
11. Kolpakov F.A. and Ananko E.A. (1999) Interactive data input into the GeneNet database. Bioinformatics, 15, 713–714.
12. Logvinenko N.S., Khlebodarova T.M., Solenov E.I., Ivanova L.N., Broude N.E., and Monastyrskaya G.S. (1991a) Hormonal regulation of Na+,K+-ATPase mRNA expression in rat kidneys during the postnatal ontogenesis. Tsitologiya, 33, 18–25.
13. Logvinenko N., Bourbouze R., Corthesy-Theulaz I., Rossier B., Bonvalet J.P., and Farman N. (1991b) Effect of aldosterone status on the expression of α ATPase isoforms in the rat kidney during post-natal development. In: Aldosterone: Fundamental Aspects. J.P. Bonvalet, Ed., Paris: INSERM, 25, 324.
14. Logvinenko N.S., Dulubova I., Fedosova N., Larsson S.H., Nairn A.C., Esmann M., Greengard P., and Aperia A. (1996) Phosphorylation by protein kinase C of the α-1 subunit of rat Na+,K+-ATPase affects its conformational equilibrium. Proc. Natl. Acad. Sci. USA, 93, 9132–9137.
15. Marver D. (1992) Regulation of Na+,K+-ATPase by aldosterone. Semin. Nephrol., 12, 56–61.
16. Schneider M., Ulsenheimer A., Christ M., and Wehling M. (1997) Nongenomic effects of aldosterone on intracellular calcium in porcine endothelial cells. Endocrinology, 138, 3410–3416.
17. Skou J.Ch. and Esmann M. (1992) The Na+,K+-ATPase. J. Bioenerg. Biomembr., 24, 249–261.
18. Verrey F., Kraehenbuhl J.P., and Rossier B.C. (1989) Aldosterone induces a rapid increase in the rate of Na+,K+-ATPase gene transcription in cultured kidney cells. Mol. Endocrinol., 3, 1369–1376.
19. Verrey F., Beron J., and Spinder B. (1996) Corticosteroid regulation of renal Na+,K+-ATPase. Miner. Electrolyte Metab., 22, 279–292.
20. Wehling M. (1995) Nongenomic aldosterone effects: The cell membrane as a specific target of mineralocorticoid action. Steroids, 60, 153–156.

DEVELOPMENT OF KNOWLEDGE BASE ON PLANT GENE EXPRESSION REGULATION

Stepanenko I.L., Goryachkovsky T.N., Ibragimova S.S., Axenovich A.V., Omelyanchuk N.A., Lavryushev S.V., Podkolodny N.L.

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

Keywords: database, gene network, gene expression regulation, transcription regulation, ontology

Introduction

Considerable recent attention has been focussed on molecular regulatory mechanisms of plant gene expression regulation. Experimental data available make possible to conclude about similar mechanisms of gene expression regulation in plant and animal cells. Plants and animals have homologous genes providing cell metabolism: genes of ribosomal proteins, translation factors, molecular chaperones; enzymes of glycolysis, amino acid biosynthesis, etc. Gene expression regulation in cells is governed by transcription factor families common for plants and animals (MADS box, homeobox, myb, leucine zipper, zinc finger). The discriminative features of plant gene regulation may be revealed at the level of an organism, during ontogenesis, and in specialized tissues and cells. Genetics has passed the way from revealing and analysis of mutations modifying ontogenesis to the understanding of mechanisms of action of proteins and their complexes during the cell differentiation, under formation of a definite gene expression pattern necessary for communication between cells and between a cell and outer environment. The next stage should include the synthesis of the stored data, description of molecular processes of gene expression regulation by using the terms and notions of plant morphogenesis. The goal of the present work is to systematize an information on molecular mechanisms of plant gene expression regulation at the level of a whole organism.

Results

Key events in gene expression regulation occur at the level of transcription. In the database TRRD (Transcription Regulatory Regions Database), an information is accumulated about structure-functional organization of gene regulatory regions in eukaryotes [Kolchanov N.A. et al., 2000]. The section PLANT-TRRD contains description of transcription regulation of more than 140 plant genes (http://wwwmgs.bionet.nsc.ru/mgs/papers/goryachkovsky/plant-trrd/).

Table 1. Informational content of PLANT GENE EXPRESSION KNOWLEDGEBASE.

Section	Description	References
PLANT-TRRD	Database on plant gene transcription	[Goryachkovsky T.N. et al., 2000a]
DATABASE ON NETWORKS OF GENES PROVIDING PLANT GROWTH AND DEVELOPMENT	Biosynthesis of storages in the process of seed maturation Degradation of storages under seed germination Plant cell response on infection by pathogens	[Goryachkovsky T.N. et al., 2000b] [Axenovich A.V. et al., 2000] [Goryachkovsky T.N. et al., 2000c]
PLANT ONTOLOGY	Plant morphology Main tissues Pathways of cell differentiation Life cycle of plants Glossary of key botanical terms	

Gene expression regulation in a cell is produced within the frames of complex systems, gene networks which include not only coordinately functioning genes but the products of their expression, phytohormones, and different external signals. The format of the database GeneNet [Kolpakov F.A. et al., 1998] is applied under description of molecular mechanisms in regulation of plant growth and development. Besides, the database contains (i) descriptions of processes (http://wwwmgs.bionet.nsc.ru/systems/Mgl/GeneNet) occurring during development and germination of seeds, (ii) interaction of plants with the other organisms (pathogens). The software program GeneNet Viewer automatically generates the scheme of a gene network on the basis of formalized data stored in the base [Kolpakov F.A. et al., 1999].

Accumulation of knowledge on molecular bases of gene expression regulation and formalized description of processes going on in the plant organism makes necessary to create a description of plant ontogenesis, which will help to trace the pathway from a gene to the process at the level of an organism. Despite enormous variety of plants, all of them are based on similar principles and have a modular construction. A typical module consists of a leave, stem, and gemma. All the plants have the similar specialized types of cells, which are organized into

functional groups of tissues: meristem or developing tissue, cover tissue, basic and conductive tissues. The knowledge base is supplied by illustrated glossary on main organs, tissues, and cells of plant organism (Figure 1).

As specific features of a plant cell may be viewed (i) an ability to bind carbon dioxide in the process of photosynthesis, due to which the plant may function in autonomous regime, and (ii) the presence of rigid cell wall, modification of which leads to formation of formalized cells. Plant ontogenesis may be viewed as alternation of stages of cell growth and differentiation. During cell differentiation, the determining factors are disposition of cell division plate and their spatial regulation. These processes are controlled by external stimuli like light, gravitation, temperature, nutritions, and phytohormones. It is of common knowledge that plant cell has a unique ability: it is totipotent, that is, under definite conditions, the differentiated cell may regenerate the entire plant.

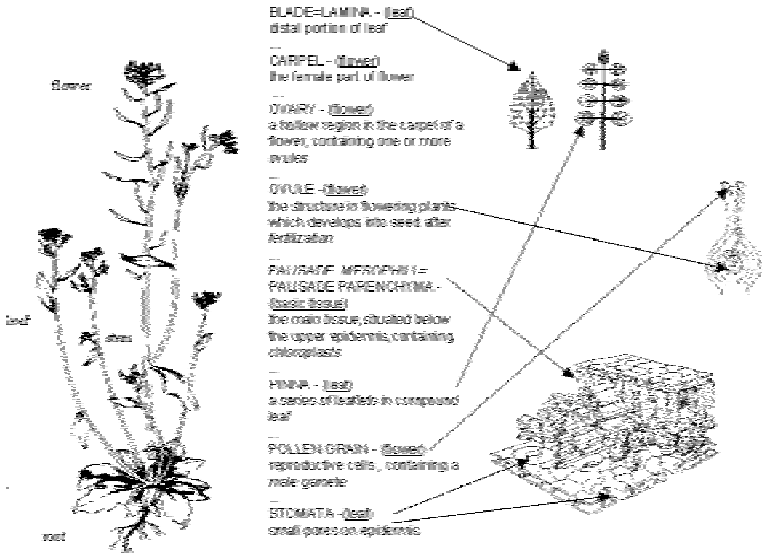


Figure 1. Plant ontology: plant morphology, main systems of tissues, organs, and glossary.

The knowledge base includes ontology of plants and experimental data on gene expression regulation stored in the databases GeneNet and TRRD (Table 1). This base is oriented to molecular biologists and theoreticians.

Conclusion

In future, by using methodology on gene network GeneNet, we plan to describe molecular mechanisms of plant ontogenesis, from fertilization to formation of an embryo and specialized types of tissues and organs in the database PLANT GENE EXPRESSION KNOWLEDGEBASE. Besides, there will be stored an information on gene networks of nitrogen fixation and photosynthesis – unique processes typical of plants. At the next stage, a description of gene networks will be done, by taking into account their hierarchical organization and spatial distribution of the processes going in a plant organism. Development of the knowledge base allows for modeling of gene network events crucial for this or that process. We plan to incorporate computer-assisted data analysis of plant gene regulatory regions and analysis of gene transcription regulation in dicots and monocots during ontogenesis, together with comparison of transcription factors and the data on structure-functional organization of regulatory regions.

Acknowledgment

This work is supported by the Russian Foundation for Basic Research (grant No 00-04-49255).

References

1. Kolchanov N.A. et al., (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. Nucleic Acids Res., **28**, 298-301.
2. Kolpakov,F.A., Ananko,E.A. (1999) Interactive data input into the GeneNet database. Bioinformatics, **15**,713-714.
3. Kolpakov,F.A., Ananko,E.A., Kolesov,G.B., Kolchanov,N.A. (1998) GeneNet: a database for gene networks and its automated visualization. Bioinformatics, **14**, 529-537.
4. Axenovich A.V., Goryachkovsky T.N., Ananko E.A., Omelyanchuk N.A., Stepanenko I.L.(2000) Gene network on storage mobilization in seed. Proceedings of BGRS'2000 (this issue), Novosibirsk.
5. Goryachkovsky T.N., Ananko E.A., Peltek S.E.(2000a) PLANT-TRRD database. Proceedings of BGRS'2000 (this issue), Novosibirsk.
6. Goryachkovsky T.N., Ananko E.A., Kolpakov F.A., Stepanenko I.L. (2000b) Seed germination in higher plants: gene networks on ontogenesis in storage tissues. Proceedings of BGRS'2000 (this issue), Novosibirsk.
7. Goryachkovsky T.N., Ananko E.A., Kolpakov F.A. (2000c) Gene network on plant interaction with pathogen organisms. Proceedings of BGRS'2000 (this issue), Novosibirsk.

FUNCTIONAL GENE NETWORKS – A DATA MANAGEMENT APPROACH FOR BIOINFORMATICS

Gabrielyan O.R.,** [1]Freytag J.C.***

Kelman Gesellschaft für Geninformation mbH, Berlin, Germany
[1]Humboldt-Universitaet zu Berlin, Institut fuer Informatik, Berlin, Germany
e-mail: freytag@informatik.hu-berlin.de
*Corresponding author

Keywords: database technology, allelic variants, splice forms, polymorphisms, gene network

Resume

Motivation:

With the breath-taking progress of public and private genome, transcriptome and proteome initiatives, computer science has to cope with several new tendencies and challenges. Gene and protein sequence collections continue to grow exponentially, and, at the same time, diversity details (polymorphisms, splice forms) as well as functional insights (disease-causing mutations, differential expression data) are being split among heterogeneous by content and format data sources. Static hyperlinks between information bits allow a researcher to navigate in the best-curated, web-based data sources and to retrieve the underlying flat files.

Nevertheless, the current state of data management remains far behind the needs of a simultaneous activation of sequence related facts for a given biocomputing purpose.

Results and discussion:

To overcome mentioned above obstacles, modern achievements in data base technology, real-world-driven data atomization and modeling, the implementation of user-defined functions into the kernel of the database, parallelized algorithms, a smart layer architecture, virtual and materialized views and other appropriate considerations are being incorporated into Kelman's high-end concept of bioinfomatics and functional genome research. These novelties ensure on principle new levels of data consistency and exploitation, and, thus, pave the way to an in-depth understanding of *gene interplay*, involving genes in all their allelic variants, transcripts in different splice-forms and post-synthetic protein processing products. Collectively, these sequence derivatives comprise a puzzling variety of molecular versions on gene product level and, therefore, require an adequate handling as well as differential investigation. The relevance of distinct molecular versions for health and illness, evolution and economy is subject to the Gene Network concept to be presented at the conference. It will be exemplified on hereditary disease research, how complex sequence-function relationships can be unraveled by using the above approach.

Availability:

Internet address of Kelman's Gene Network concept is http://www.kelman.de

GENE NETWORK ON PLANT INTERACTION WITH PATHOGEN ORGANISMS

Goryachkovsky T.N., Ananko E.A., Kolpakov F.A.

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

Keywords: plant pathogen, pathogenesis-related genes, hypersensitive response, gene networks, databases

Resume

Motivation:

Higher plants are equipped with a set of mechanisms protecting them from diseases caused by pathogen bacteria, fungi or viruses. Studying of protective mechanisms providing disease resistance in plants is a promising source in perspective developing of new varieties and accessions of plants, in synthesis of artificial molecular-genetic constructs with the pre-determined characteristics of gene expression. During the recent period, molecular mechanisms providing recognition of a pathogen, signal transduction, and activation of protective systems are being intensively studied. The convenient tools for accumulation and systematization of the ever-growing bulk of information in this field are the databases TRRD and GeneNet, which were developed in our laboratory [Kolchanov N.A. et al., 2000, Kolpakov F.A. et al., 1998].

Results:

The TRRD database stores an information on regulatory gene regions responsible for the plant reaction in a response to pathogen infection. By applying the tools of GeneNet database, the gene networks on interaction between plant and pathogenic microorganisms are described. These gene networks contain an information about elementary structures and events of this process. In the current TRRD version, an information is accumulated on regulatory regions of 20 pathogen-induced genes. In GeneNet database, the elementary events occurring during plant interaction with a pathogen are described for 5 different plant species.

Availability:

The gene network “INTERACTION BETWEEN PLANT AND PATHOGEN” is a section of the GeneNet database and it is available via the Internet at http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/

The corresponding section of the TRRD database can be found via

http://wwwmgs.bionet.nsc.ru/mgs/papers/goryachkovsky/plant-trrd/

Introduction

During the contact between plant and pathogenic microorganism, a particular chain of events is produced in the plant organism. The interaction between plant and pathogen may develop by two ways given below.

1. The plant is provided by a receptor that interacts with bacterial protein. As a result, quick protective reaction is being developed. In such a situation, the bacteria is called avirulent for a given plant genotype [Piffanelli P. et al., 1999, Martin G.B., 1999].
2. The proteins of the pathogenic organism are virulent for the given plant genotype. The plant is affected by the pathogen, whereas protective mechanisms are being activated more slowly [Maleck K and Lawton K., 1998].

In both cases, with the start of pathogenesis gene transcription, the cell walls strengthen. Then in the place of pathogen penetration, the active forms of oxygen are formed, causing the death of infected cells.

Results

By graphical representation (visualization) of information stored in GeneNet database, the gene networks were constructed, which are functioning in plant after induction by pathogen. The format of GeneNet database enables to describe protein localization within the cell, an extent of the protein multimerization (i.e., monomer, dimer or multimer), and the state of a protein (phosphorilated or non-phosphorylated, active or inactive) [Kolpakov F.A. et al., 1998]. Gene networks activated in the course of plant interaction with pathogenic organisms are referred to the networks providing response to environmental conditions. Let us consider in details the chain of events occurring during plant interaction with a pathogen (Fig.1).

At the present moment, the best studied are the molecular mechanisms providing hypersensitive response (HR) [Halterman D.A. and Martin G.B., 1997]. In this case, the plant receptor interacts with the pathogen molecule. In order such interaction could occur, the plant and bacteria of a certain genotype should meet, i.e., a bacteria carrying the avirulence gene (avr) interacts with a plant, which has the corresponding R-gene. Such process is called an incompatible combination and leads to quick progressing of events, or to hypersensitive response. Receptors activate the passes of signal transduction and launch several protective systems. In the place of pathogen penetration, the strong oxidants are being synthesized such as H_2O_2, $O_2\bullet -$, $OH\bullet$. Then the oxidative burst is being developed, followed by the death of infected cells according to the mechanism similar to apoptosis known in vertebrates [Lam E. et al., 1999]. Hypersensitive response is being developed in the place of pathogen penetration into the neighboring cells. Rapidly developing local process produces signal molecules spreading along vascular system of a plant.

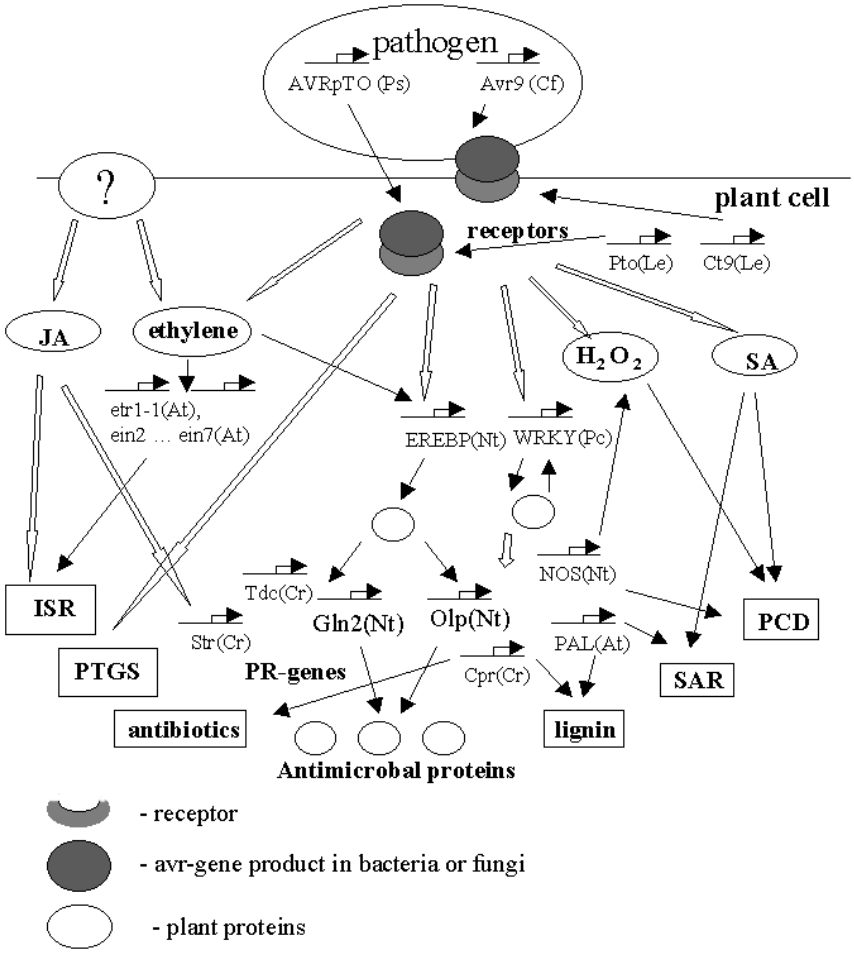


Figure 1. The scheme of molecular events occurring during pathogenesis of plants. SA - salicylic acid; JA - jasmonic acid; H2O2 - reactive oxygen species; PR - pathogenesis-related genes; SAR -systemic acquired resistance; ISR - induced systemic resistance; PCD- programmed cell death; PTGC - post-transcriptional gene silencing. By thin arrows are given direct involvement of a substance of gene product in a reaction. Empty arrows denote the events that need additional elements for signal transduction.

In dependence of local events, the set of signal molecules is being organized, which in turn, forms this or that generalized response. In the whole plant organism, the pathogenesis-related genes (PR-genes) are activated, the cell walls strengthen, and the plant accumulates some amount of protective substances, which are more effective in the struggle with this definite pathogenic form. In the plant cells, the salicylic acid (SA) is produced in considerable amounts and causes activation of SA-induced genes. The integrity of these events is named as systemic acquired resistance (SAR) [Mittler R. et al., 1999]. The other systemic, or referring to the whole organism, response to pathogen infection is an induced systemic resistance (ISR). Its differs from SAR by activation of some differing set of pathogenesis genes and by other ways of signal transduction (without participance of salicylic acid) [Pieterse C.M.J. et al., 1996].

In addition to immunity specific to pathogen, the general protective mechanisms are launched and, the plant becomes more resistant to other diseases after meeting the aviral pathogen [Zhu Q. et al., 1996].

The plant is not always supplied with receptors to the proteins of "attacking" bacteria or fungi. In this situation, the pathogen is called viral for given plant genotype; and the pair plant-pathogen is compatible. In this case, pathogen molecules are non-specific elicitors, which are non-specific substances causing pathogenesis. The ways of obtaining the signal from non-specific elicitors are still unknown. Various external stimuli (wound, non-specific elicitors) activate protein kinases and genes of signal molecules biosynthesis. In the course of signal transduction, the syntheses of JA, NO, H_2O_2, SA, and ethylene are produced. The processes are being activated that are well-known in animals: proteinkinase cascade, polyubiquitine-dependent protein degradation, etc. The signal transduction paths frequently intercross. For example, the gene of one of the key enzymes, PAL (phenylalanine ammonia-lyase), is activated not only in the course of hypersensitive response, but also in response to various external stimuli [Dixon R.A. and Paiva N.L. 1995, Mauch-Mani B. et al., 1996]. PAL takes part in the synthesis of SA, phytoalexins, and lignin monomers.

The latter, both in hypersensitive response and non-specific pathogenesis induction, activate pathogenesis-related (PR) genes. To these genes are referring those encoding enzymes of protective substances biosynthesis, chitinases acting in degradation of fungi cell wall, enzymes of lignin biosynthesis, which is a component of plant cell wall, etc.

The positive and negative feedbacks stabilize the parameters of gene networks. For example, WRKY, transcription factor proceeding in PR-genes activation is regulated according to the principle of positive feedback, because it has the binding site in its own promoter. Along with H_2O_2 and salicylic acid biosynthesis, the events are developing, which lead to the death of infected cells. The cell dies together with the source of signaling molecules. Hence, the gene network terminates its functioning by the mechanism of the negative feedback.

An analysis of information accumulated in GeneNet and TRRD databases makes able to separate three groups of genes acting in protection of plant from virulent and avirulent pathogens (Table 1).

Table 1. Classification of genes participating in the work of gene networks, in response to induction by a pathogen or other damaging agents.

Species of a plant	R-genes (receptors)	Genes of the primary response	PR-genes
Tomato	Pto, Cf-4, Cf-9, Cf-2, Cf-5,		
Tobacco	N	ACC synthase, EFE, NO synthase, Ubi.U4	osmotin, β-1,3-glucanase, Chn48, Chn50, PRB-1b, PRB-1a, PRB-1b, str246C
Barley	Mlo		lipid transfer protein, Lox1, blt4.9, Ltp4.3, lipoxygenase 1,
Maize		(Lls)1	lipid transfer protein
Sugar beet	$Hs1^{pro-1}$		
Rice	Xa21, Xa1, Pi-ta,		
Flax	L,		
Arabidopsis	Rpm1, Rps5,	Nim1/npr1/Sai1, Lsd1, Acs6	Adh
Horseradish			prxC2
Parsley		PAL, 4CL1, 4CL2	PR1, PR1-1, PR1-2
Periwinkle			CPR

To the first group of genes, we refer receptors of plants interacting with the avr-gene product of bacteria or fungi. In the literature, it is accepted to denote as R-genes the receptors and some genes involved in signal transduction. Based on the recent studies of molecular mechanisms of plant response to pathogenic infection, it seems to us reasonable to unify all genes of signal transduction paths into a single group of primary response. To these group are referred the genes both transferring the signal from R-receptors and from still unknown receptors of non-specific elicitors, genes of biosynthesis of signal molecules, protein kinases, etc. To the third group we suggest to refer the genes of the final stage in a chain, which is activated in response to pathogenic infection, or the so-called PR-genes (pathogenesis-related genes).

Discussion

Contemporary achievements in the field of plant gene engineering put the foundation for rapid growth of technologies for producing the plants with pre-determined characteristics. At this stage, for the most actual one, we consider an approach based on gene network concept, which states that all the processes in an organism are provided by coordinated expression of various groups of genes [Kolpakov F.A. et al., 1998]. Each group of this type makes a foundation for a gene network that provides the fulfillment by a cell or by an organism of a definite function or the function that is responsible for production of a concrete trait. Under a gene network, here we mean the integrity of coordinately expressed genes, their protein products, and interactions between them. As the elements of gene networks are viewed the low-molecular substances, external signals, hormones, metabolic products, etc. Regulation of gene network functioning is not restricted only by the level of

transcription. Translation, splicing, post-translational protein degradation, active membrane transport, etc may also produce the regulation. The switch between gene networks takes place as a rule in accordance with the stage of ontogenesis or under the action of external signals.

Such approach enables to model the behavior of heterogeneous genes in the organism of transgenic plants, to make account of gene networks stabilizing parameters, which were destroyed by transferred gene. The practical application of the gene network concept opens new possibilities of constructing the plants with new genotypes supplied by increased disease resistance or other valuable productive traits.

Acknowledgements

This work was supported by the Russian Foundation for Basic Research (grants Nos. 00-04-49229, 00-04-49255, 00-07-90337). The authors are grateful to G. Orlova for translation of the manuscript into English.

References

1. Dixon,R.A, Paiva,N.L. (1995) Stress-induced phenylpropanoid metabolism. Plant Cell, 7, 1085-1097.
2. Halterman,D.A. and Martin,G.B. (1997) Signal recognition and transduction involved in plant disease resistance. Essays Biochem., 32, 87-99. Review.
3. Kolchanov,N.A., Podkolodnaya,O.A., Ananko,E.A., Ignatieva,E.V., Stepanenko,I.L., Kel-Margoulis,O.V., Kel,A.E., Merkulova,T.I., Goryachkovskaya,T.N., Busygina,T.V., Kolpakov,F.A., Podkolodny,N.L., Naumochkin,A.N., Korostishevskaya,I.M., Romashchenko,A.G., Overton,G.C. Transcription regulatory regions database (TRRD): its status in 2000. (2000) Nucleic Acids Res., 28(1), 298-301
4. Kolpakov,F.A., Ananko,E.A., Kolesov,G.B. and Kolchanov,N.A. (1998) GeneNet: a database for gene networks and its automated visualization. Bioinformatics,14(6), 529-537.
5. Lam,E., Pontier,D. and del Pozo,O. (1999) Die and let live - programmed cell death in plants.Current Opinion in Plant Biology, 2(6):502-507.
6. Mauch-Mani,B. and Slusarenko,A,J. (1996) Production of salicylic acid precursors is a major function of phenylalanine ammonia-lyase in the resistance of Arabidopsis to Peronospora parasitica. Plant Cell, 8, 203-212.
7. Maleck,K and Lawton,K., (1998) Plant strategies for resistance to pathogens. Curr.Opin.Plant.Biol., 9, 208-213.
8. Martin,G.B., (1999) Functional analysis of plant disease resistance genes and their downstream effectors. Curr.Opin.Plant.Biol., 2, 273-279.
9. Mittler,R., Lam,E., Shulaev,V. and Cohen,M. (1999) Signals controlling the expression of cytosolic ascorbate peroxidase during pathogen-induced programmed cell death in tobacco. Plant.Mol.Biol., 39(5),1025-1035.
10. Pieterse,C.M.J., van Wees,S.C.M., Hoffland,E., van Pelt J.A. and van Loon,L.C. (1996) Systemic resistance in Arabidopsis induced by biocontrol bacteria is independent of salicylic acid accumulation and pathogenesis gene expression. Plant Cell, 8, 1225-1237.
11. Piffanelli,P., Devoto,A. and Schulze-Lefert,P. (1999) Defence signalling in cereals. Curr.Opin.Plant.Biol., 2, 295-300.
12. Zhu,Q., Droge-Laser,W., Dixon,R.A. and Lamb,C. (1999) Transcriptional activation of plant defense genes. Curr.Opin.Genet.Dev., 6(5), 624-30

PUMA/WIT -- A FAMILY OF INTEGRATED SYSTEMS FOR GENETIC SEQUENCE ANALYSIS AND METABOLIC RECONSTRUCTIONS

Overbeek R., Selkov E., Pusch G., D'Souza M., *Maltsev N.

Argonne National Laboratory, Argonne, Illinois, USA
e-mail: maltsev@mcs.anl.gov
*Corresponding author

Keywords: metabolic pathways, metabolic reconstructions, sequence analysis, regulatory networks, database

Resume

Recent years have witnessed a rapid accumulation of sequence data and data related to the physiology and biochemistry of organisms. Providing a functional context for this data is vital to understanding of complex biological systems. Adequate representation of metabolism is one of the keys to design of such context.

The Computational Biology group at the Mathematics and Computer Science Division of Argonne National Laboratory has designed and implemented a family of PUMA/WIT integrated systems for sequence analysis and metabolic reconstruction from the sequence data based on EMP/MPW collection of metabolic pathways [1]. Table 1 presents a brief summary of the major features of these systems.

Table 1. A brief summary of the major features of PUMA/WIT systems.

Year	System	Description	Authors	Comments
1995	PUMA	Integrated system providing access to sequence information within a context of metabolic pathways from the MPW collection, general functional overview, and phylogenetic tree. Provided links to public sequence databases, multiple sequence alignments alignments, etc.	Ross Overbeek, E. Selkov, N. Maltsev, T. Gaasterland	[2]
1996	WIT	Integrated system that allowed interactive genetic sequence analysis and metabolic reconstruction. Provided access to a large variety of genetic sequence analysis tools and allowed user driven annotations..	Ross Overbeek, E. Selkov, N. Maltsev, N. Larsen	[3]
1997-current	WIT2	Current version of WIT2 with much-improved user interface in comparison with WIT; contains 38 completely and almost completely sequenced genomes and provides access to the originally developed set of tools: orhtolog clusters, conserved chromosomal gene clusters [19], chromosomal maps, etc.	Ross Overbeek, E. Selkov, N. Maltsev, G. Pusch, M. D'Souza, N. Larsen	[4], [5]
1999-current	PUMA2	An environment for comparative analysis of metabolic subsystems and automated reconstruction of metabolism of microbial consortia and individual organisms from sequence data.	Natalia Maltsev, Mark D'Souza	[6]

The WIT2 system

The WIT2 system (http://wit.mcs.anl.gov/WIT2) supports genetic sequence and comparative analysis of sequenced genomes as well as metabolic reconstructions from the sequence data. It now contains data from 39 completely and almost completely sequenced genomes. It provides access to thoroughly annotated genomes within a framework of metabolic reconstructions, connected to the sequence data; data on regulatory patterns, protein alignments and phylogenetic trees; as well as data on gene clusters, potential operons and functional domains. We believe that the parallel analysis of a large number of phylogenetically diverse genomes simultaneously can add a great deal to our understanding of the higher level functional subsystems and major physiological designs.

The purpose of WIT2 project is to:

- utilize the advantages of comparative analysis of the genomes for the genetic sequence analysis develop reconstructions of the metabolic network for sequenced genomes;
- develop approaches to understanding of sensory transduction and regulatory networks;
- develop automated tools for high throughput genetic sequence analysis and metabolic reconstructions;
- develop a framework for interaction with the biological community and produce a set of conjectures; about the possible functions of the genes and modes of function of biological subsystems to be tested in the wetlab.

WIT provides an environment for comparative analysis of the genomes. Variety of genetic sequence analysis tools developed by the other groups (such as Blast, Fasta, ProDom, COGs, Pfam, TMpred etc) is accessable within the WIT2 framework. We have also developed a number of original methods and approaches to exploit the advantages of comparative analysis of multiple genomes. These methods include the following:

- Development of protein clusters that allow to increase dramatically the speed and precision of assignments of functions to the genes in newly sequenced genomes. WIT2 team currently maintains a set of ortholog clusters that includes the sequences from 39 genomes available in WIT;
- Use of contiguity on the chromosome to predict functional coupling [7];
  WIT2 team has developed an original method for prediction of functional coupling between the genes based on presence of conserved chromosomal gene clusters. This technique, which depends on the availability of a relatively large number of genomes, offers the opportunity of capturing significant clues to function for many genes of unknown function as well as provides insight into composition of the functional subsystems;
- Use of spreadsheet to predict pathways
  The current version of WIT2 contains a spreadsheet that allows researchers to quickly estimate which organisms of WIT's 39 have a given pathway or a fragment of a pathway;
- General Functional Overview
  Biological systems are characterized by an incredible metabolic and physiological versatility. The same function (for example, utilization of glucose) could be performed by different organisms and even by the same organism under the different conditions in a number of different ways, employing different enzymes, regulation circuits and intermediate products. The idea behind the organization of the available biological data in a context of a General Functional Overview is to bring together similar biological functions, performed in the different organisms and under different condition and to provide a functional context for a rapidly emerging body of the sequence data.

Metabolic Reconstructions.The first and primary goal of WIT2 is to develop, maintain, and distribute detailed metabolic reconstructions for a broad variety of phylogenetically diverse sequenced organisms [3]. The utility of metabolic reconstructions is now widely recognized---both for the development of more accurate function assignments for genes and their products and, more important, as a first step toward imposing a functional overview of the cell and characterizing the dynamic behavior of such systems. These reconstructions will represent a major advance in our understanding of the metabolism of the organisms and will benefit research in such fields as biotechnology, bioremediation, and medicine. Metabolic reconstructions developed for prokaryotes will provide a significant basis for our understanding of how this metabolism relates to that of eukaryotes.

WIT2 currently contains comprehensive metabolic reconstructions for more than 25 organisms. Key to success in this effort is to establish active collaborations with the experts in metabolism of each of the sequenced species.

SENTRA -- A Database of Signal Transduction Proteins

In collaboration with Dr. M. Romine (Pacific Northwest National Laboratory) we are developing SENTRA (http://wit.mcs.anl.gov/WIT2/Sentra) [8]-- a WIT-based database of proteins associated with microbial signal transduction. This database currently includes the classical two component signal transduction pathway proteins and methyl-accepting chemotaxis proteins from 34 complete and almost completely sequenced prokaryotic genomes, as well as sequences from 243 organisms available in public databases (SwissProt and EMBL).

Every entry in Sentra database (from WIT genomes) is annotated with the information on functional domains (Pfam) and motifs, transmembrane domains, conserved chromosomal gene clusters, relevant protein family.

The PUMA2 system

Recently we have started to develop PUMA2 [6] -- an environment for comparative analysis of metabolic subsystems and automated reconstruction of metabolism of microbial consortia and individual organisms from sequence data.

PUMA2 intends to:

- Allow comparative analysis of the metabolic subsystems in different organisms;
- Provide a framework for the automated reconstruction of the metabolism of microbial consortia and individual species;
- Provide a framework for representation of the expression data;

A working prototype of PUMA2 system can be viewed at http://puma.mcs.anl.gov/PUMA2. Analyses in PUMA2 will be based on a collection of metabolic modules connected to sequence data. Every "metabolic module" represents a set of enzymes involved in particular physiological process (e.g., glycolysis, methanogenesis, nitrogen assimilation) with all known variations in different organisms, presented in the form of a metabolic diagram. Every module is connected to the other modules through common intermediates and components, thus constituting an architecture of a metabolic network.

We will annotate each metabolic module with the information on the signature enzymes or combination of the enzymes. These signatures uniquely characterize a given module, and their presence in a given organism will provide an unambiguous evidence of occurrence of this module in an organism or microbial community. The results of such analyses are presented in a graphical form based on hierarchical representation of the functional subsystems and annotated with sequence data and literature information.

Use of PUMA2 environment will permit the following applications:

1. Identification of signature genes for characterization of biological communities or single isolates
2. Development of a tool kit for visualization and annotation of user defined pathways 3.Identification of unculturable organisms
3. Prediction of properties of ecological niches based on microbial physiology

We believe that the integration of a large number of phylogenetically diverse genomes in the WIT/PUMA systems could assist the understanding of the physiology of complex biological systems.

References

1. Selkov Jr., E., Grechkin, Y., Mikhailova, N., and Selkov, E. MPW: The Metabolic Pathways Database. Nucleic Acids Res. 26(1), pp. 43-45, 1998.
2. Overbeek, R., Larsen, N., Smith, W., Maltsev, N., and Selkov, E. Representation of Function: The Next Step. Gene 191(1), pp. GC1-GC9, 1997.
3. Selkov, E., Maltsev, N., Olsen, G.J., Overbeek, R., and Whitman, W. B. A Reconstruction of the Metabolism of Methanococcus jannaschii from Sequence Data. Gene 197(1-2), pp. GC11-GC26, 1997.
4. Overbeek, R., Larsen, N., Maltsev, N., Pusch, G. D., and Selkov, E. In: Stan Letovsky (ed.) Molecular Biology Databases, Kluwer Academic Pub., pp. 158-167, 1999.
5. Overbeek, R., Larsen, N., Pusch, G., D'Souza, M., Selkov Jr., E., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. WIT -- Integrated System for High-throughput Genetic Sequence Analysis and Metabolic Reconstructions. Nucleic Acids Res 2000 Jan 1;28(1):123-125.
6. M. D'Souza, N. Maltsev. PUMA2 -- an environment for comparative analysis of metabolic subsystems and automated reconstruction of metabolism of microbial consortia and individual organisms from sequence data. Technical memorandum ANL/MCS-TM-240.
7. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G., and Maltsev, N. The Use of Gene Clusters to Infer Functional Coupling. Procceedings of the National Academy of Science U.S.A., 96(6), p. 2896, 1999.
8. D'Souza M, Romine MF, Maltsev N. SENTRA, a database of signal transduction proteins. Nucleic Acids Res 2000 Jan 1;28(1):335-336.

LATENT PHENOTYPE AS AN ADAPTATION RESERVE: A SIMPLEST MODEL OF CELL EVOLUTION

Likhoshvai V.A., Matushkin Yu.G.

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

Keywords: evolution, genotype, phenotype, mathematical model, computer analysis

Resume

Motivation:

Optimization of the efforts spent on cell reproduction is a factor capable of imparting evolutionary trend to the changes in the intracellular parameters. It is necessary to study theoretically the pattern of evolutionary effect of this factor on biological systems.

Results:

A simplest mathematical model of cell evolution is constructed and studied basing on the hypothesis on existence of the evolutionary trend to optimize the consumption of unified external energy resources. Analysis of the model demonstrates that the rates of intracellular processes change with time in such an interactive manner that the cell at a certain moment of its evolution acquires the possibility of existing in at least two stable states. The cell, having one and the same genotype as well as the constants of metabolic reaction rates, can still have two essentially distinct phenotypes.

Introduction

It is supposed conventionally that the state of an organism is determined exclusively by its genotype and the environmental conditions. The possibility of an organism to exist in different states is always associated with the genetically fixed regulatory mechanisms, manifesting themselves under certain conditions. However, there are systems whose alternative states cannot be derived from the structures of their subsystems. For example, a simplest molecular genetic trigger can function in two alternative states; this is the feature of the system on the whole, whereas no its subsystem alone possesses this feature. The question arises on whether autonomous biological systems (for example, cells) in the course of slow and gradual adaptation to the environmental conditions may became more complex through acquiring the possibility of existing in alternative states that cannot be reduced to genetic mechanisms and rates of metabolic reactions. To study this problem from theoretical standpoint, a simplest mathematical model of evolution of a single cell reproduction cycle is constructed in this work. We proceed from the assumption that the biological systems tent to reach homeostasis. Within our simulation method, we consider the cell homeostasis as a nontrivial stationary point of the set of differential equations describing this cell. It is demonstrated that most general organization patterns of unicellular organisms—occurrence of translation machinery, template-based protein synthesis, and reproduction through division—result in complex nonlinear relations between different parts of the cell as well as between the cell and the environment. This provides for a range of parameter values where the cell may have several alternative homeostatic states. Computer simulation of the evolutionary development on the hypothesis of the selection for increased efficacy of external energy consumption demonstrates that this range of parameters is the region of evolutionary attraction.

Model

The model of single cycle of cell reproduction considers four types of intracellular components: ribosomes (R), total protein (P), the mRNA coding for ribosomal protein (F_r), and the mRNA coding for the total protein (F_p). It is assumed that the cell consumes for its vital functions a pooled resource (Z), which is spent on syntheses of mRNA and proteins. The cell draws this resource from the environment, whereto it arrives at a constant rate. It is considered that the span of a single cell reproduction cycle is equal to the time necessary to reach the threshold concentration ($\overline{P}$) of the total protein. Syntheses of total proteins $P_r^{(i)}$ and $P^{(i)}$ is described by the following reactions:

$$R + F_r \xrightarrow{b_r z^{(i)}} F_r^* \xrightarrow{3k_e z^{(i)}/n_s} F_r + R_r$$
$$R_r \xrightarrow{k_e z^{(i)}/(n_r - n_s/3)} R + P_r^{(i)} \qquad (1)$$

$$R + F_p \xrightarrow{b_p z^{(i)}} F_p^* \xrightarrow{3k_e z^{(i)}/n_s} F_p + R_p$$
$$R_p \xrightarrow{k_e z^{(i)}/(n_p - n_s/3)} R + P^{(i)} \quad (2)$$

Let us designate the homeostatic parameters of the maternal cell with the superscript (i); of the daughter cell, (i+1). Let us calculate the equilibrium concentration of the pooled resource for the maternal cell, assuming that it is in equilibrium and considering the equilibrium ratio for ribosomes, mRNA, and the pooled resource:

$$z^{(i)} = s_z / (d_z + \varepsilon_n (m_r F_r^0 + m_p F_p^0)\overline{v_p^{(i)}} / \overline{P} + \varepsilon_a (\overline{v_r^{(i)}} + \overline{v_p^{(i)}}),$$
$$\overline{v_r^{(i)}} = F_r^0 / (n_s / (3k_e) + 1/(b_r R)), \overline{v_p^{(i)}} = F_p^0 / (n_s / (3k_e) + 1/(b_p R)) \quad (3)$$

where R is the unique positive solution of the ribosome equilibrium equation:

$$R + (n_r / k_e)\overline{v_r^{(i)}} + (n_p / k_e)\overline{v_p^{(i)}} = R_0^{(i)} \quad (4)$$

and the span of (i)th cell cycle is

$$T^{(i)} = \overline{P} / (\overline{v_p^{(i)}} z^{(i)}) \quad (5)$$

Ignoring the degradation processes, let us calculate the number of ribosomes produced in the maternal cell over the time $T^{(i)}$:

$$\overline{R_0^{(i)}} = \overline{v_r^{(i)}} z^{(i)} T^{(i)} / P_{rib} \quad (6)$$

Then, the daughter cell will get the following number of ribosomes, symmetrical division provided:

$$R_0^{(i+1)} = (R_0^{(i)} + \overline{R_0^{(i)}}) / 2 \quad (7)$$

To simplify the model, let us assume that the mRNA concentrations in the daughter cell are equal to those in the maternal cell, if no mutation occurred. Thus, specifying the values of constants for the initial maternal cell (zero generation cell) and concentrations of ribosomes and mRNA, we can calculate the homeostatic parameters ($T^{(i)}$, $Z^{(i)}$, and $R_0^{(i)}$) for any its progenies (cells of (i)th generation) using equations (3)–(7). See table for description of the parameters and variables of the model.

Table. Parameters of the model.

No.	Symbol	Numerical value	Characteristics
1	F_r^0	1,018 pcs./cell[1,2,5]	Total concentration of the mRNA coding for ribosomal proteins
2	F_p^0	510 pcs./cell[1,2,5]	Total concentration of the mRNA coding for total protein
3	R_0	8,340 pcs./cell[2,3,5]	Total concentration of ribosomes
4	P_{rib}	54 proteins/ribosome[4]	Number of proteins in ribosome
5	$\overline{P}$	2,000,000 pcs./cell[4]	Threshold amount of total protein, determining the span of cell cycle
	T	3,433 sec[3,5]	Span of single cell reproduction cycle
6	Z	1.87 arbitrary units/cell[3,5]	Total concentration of the pooled resource
7	n_r	150 codons[4]	Length of the total ribosomal protein in amino acid residues
8	n_p	250 codons[4]	Length of the total protein in amino acid residues
9	m_r	500 nucleotides[4]	Length of F_r mRNA in nucleotides
10	m_p	800 nucleotides[4]	Length of F_p mRNA in nucleotides
11	n_s	45 nucleotides[4]	Steric size of ribosome
12	k_e	15 codons[4]	Averaged constant of elongation rate
13	b_r	3×10^{-5} initiations/sec/ribosome[1,5]	Constant of F_r mRNA translation initiation rate
14	b_p	6×10^{-4} initiations/sec/ribosome[1,5]	Constant of F_p mRNA translation initiation rate
15	ε_n	$\varepsilon_n=3\times10^{-4}$ arbitrary units of pooled resource/nucleotides	Relative energy value of nucleotides
16	ε_a	$\varepsilon_a=10^{-4}$ arbitrary units of pooled resource/amino acid residues[4]	Relative energy value of amino acid residues
17	s_z	s_z=20 arbitrary units of pooled resource/sec[4]	Constant of pooled resource input rate
18	d_z	d_z=5 arbitrary units of pooled resource/sec[4]	Constant of pooled resource degradation rate

[1]May be changed by random mutations; [2]is specified for the cell of (0) generation; [3]is calculated for the cell of (i)th generation; [4]is specified for the cell of (0) generation and remains constant at all the stages of model function; [5]the values indicated were obtained after parameter evolutions (at the parameter values given, the cell position in figure corresponds to occurrence of two homeostatic states—two phenotypes).

It is assumed that random mutations occur during the reproduction. They may change the mRNA concentrations, F_r and F_p, and the rates of translation initiation, b_r and b_p. The rest parameters are considered as fixed and remain constant from generation to generation. Mutations are assumed rare events, so that a

considerable number of generations passes between two successive mutations. Therefore, it is reasonable to assume that homeostatic characteristics of remote progenies of the mutated cell reflect more adequately the evolutionary significance of a mutation. To calculate the homeostatic parameters of remote progenies, we used the limit cases of equations (3)–(7) obtained through making *i* approach infinity (formally, the limit cases are obtained through canceling (i) and (i+1) indices).

The elementary step of evolution is simulated as follows. A random-number generator is used to select a parameter from F_r, F_p, b_r, and b_p; the value of its changing is selected basing on a distribution specified in advance; and the function parameters of a remote progeny of the "mutated" cell are calculated. If a "mutation" decreases the division span, the mutation is fixed; in the opposite case, rejected. The other variant: a mutation is fixed as a result of an arbitrary event with a probability proportional to the difference between the spans required for divisions of remote progenies of the "mutant" and "intact" cells. Thus, evolutionary trajectories of the model movement in the space of the parameters changed (F_r, F_p, b_r, and b_p) were calculated at different specified values of the constant parameters (n_r, n_p, m_r,, k_e, b_r, b_p, p_{rib}, P, ε_n, ε_a, s_z, and d_z).

Results and discussion

In the general case, the number of different limit equilibrium variants of homeostasis wherein the cell may exist exceeds 1 and is directly connected with the positive roots of the equation

$$R^3 - a_2R^2 + a_1R - a_0 = 0, \qquad (8)$$

which has at least one positive root at any combinations of the parameter values.

If the model parameter values (table) do not satisfy the following inequalities:

$$a_1 > 0, \qquad a_2 > 0, \qquad a_2^2 \geq 3a_1,$$
$$(a_2 + \sqrt{a_2^2 - 3a_1})^2(a_2 - 2\sqrt{a_2^2 - 3a_1}) \leq 27a_0 \leq (a_2 - \sqrt{a_2^2 - 3a_1})^2(a_2 + 2\sqrt{a_2^2 - 3a_1}), \qquad (9)$$

where $a_0 = (F_rP/F_p/P_{rib}) \times (k_e/n_s/b_r)$,

$a_1 = k_e^2/n_s^2/b_r/b_p + F_rn_rk_e/n_s^2/b_p + F_pn_pk_e/n_s^2/b_r - (F_rP/F_p/P_{rib}) \times (k_e/n_s/b_p)$, and

$a_2 = F_rP/F_p/P_{rib} - (k_e/n_s/b_r + k_e/n_s/b_p + F_rn_r/n_s + F_pn_p/n_s)$,

then only one limit variant for realization of intracellular metabolism exists, to which all the lines of successive cell generations converge independently of the specified non-zero initial value of ribosome concentration in the zero generation cell.

In the opposite case, the equation has at least three positive roots (possibly, multiple), corresponding to two stable and one unstable variants of limit equilibrium homeostasis of the model. In this case, the stable states correspond to the least and biggest roots of equation (3), while unstable, to the medium. If two roots of equation (3) are multiple, the corresponding state is semistable. For short, let us designate the set of parameter values unsatisfying the conditions (2) as D_1; the range of parameter values satisfying the conditions (2) as D_2.

Numeric calculations demonstrate that the volume of D_2 region amounts to an insufficient part of the D_1 region value (the ratio of D_2 to D_1 volumes amounts to 1.93×10^{-4} at the parameter values shown in table), specification of feasible values of the model constant parameters provided. Therefore, the probability of falling randomly into D_2 region is very low. Computer simulation of the evolutionary process has demonstrated that the evolutionary trajectories in the space of parameters F_r, F_p, b_r, and b_p go from D_1 into D_2 region; that is, D_2 is the region of evolutionary attraction. An example of evolutionary trajectory of cell moving from D_1 to D_2 region is shown in figure. While the evolution proceeds in D_1 region, the genotypic identity of cells necessitates their phenotypic identity (curve 1 in figure) and the diversity of the cells within the population falls within the reaction norm. The situation changes basically in D_2 region: the cells can realize their genetic program according to two alternative routes (curves 1 and 2 in figure). Both variants of homeostasis are feasible, and the cells while dividing can pass from one to the other. However, only one mode is optimal—the phenotype displaying shorter cell division span T—and the cells found themselves in the alternative state will with time be removed from the population, as they are not competitive compared with the "optimal" cells. In this sense, the nonoptimal phenotype may be considered latent. Thus, the population evolving in D_2 region contains mainly the cells with optimal phenotype, and only a small fraction of the cells develops according to the latent mode. Therefore, the evolution fixes mainly the mutations improving the function characteristics of the optimal phenotype (evolution follows curve 1). However, the function parameters of the latent phenotype are also changing (curve 2 in figure is not constant).

The transition from D_1 region, with a relatively simple behavior, to D_2 region, with a more complex behavior, during the evolution is actually a latent self-complication of the system, since the structure of the system remains the same (no new mechanisms emerge), while the set of potential variants of realizing the structurally fixed genetic program under identical conditions widens and selection of an alternative variants is randomized.

The calculations also demonstrate that the course of evolution in D_2 region tends to increase the probability of a spontaneous transition of a cell from the optimal to the latent phenotype (the separatrix in figure withdraws from curve 2 and approaches curve 1). Hence, the evolutionary value of the latent phenotype increases with time. The quantitative parameters of the latent phenotype function differ from those of the optimal phenotype; therefore, the cells that are in different states would evolve in a different way, presenting additional source of complications. However, occurrence of several variants for realizing the genotype increases the evolutionary stability of the system. For example, a drastic change in the habitat may change the parameters of the cell–environment interaction in such a way that the latent phenotype would appear more adapted to the new conditions. In this case, a saltatory exchange of the dominant phenotype will take place. However, a new latent phenotype will inevitably emerge in case this drastic change is not followed by change in the environmental conditions for an extended period.

The logical scheme used may be also applied to biological systems of a higher organizational level.

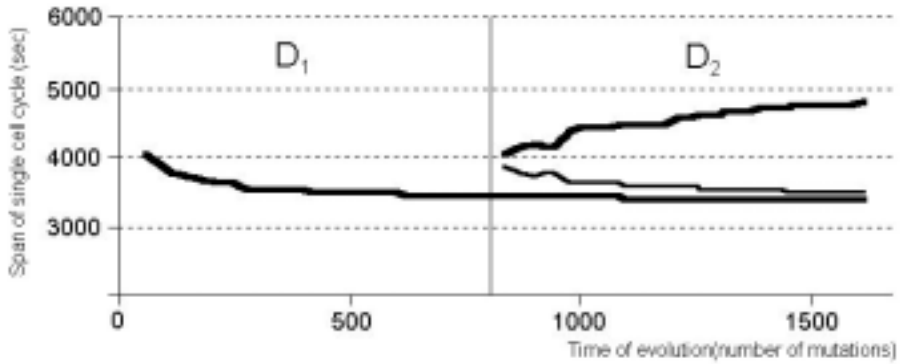


Figure. Change in the cell division span in the course of evolution (calculations according to the model; parameter values are listed in table): *(1) optimal phenotype; (2) latent phenotype; and (3) separatrix; the vertical line is the boundary between D_1 and D_2 regions.*

Acknowledgements

The work was supported by the grant No. 106 of the State R&D Program *Human Genome* and Integrative Project *Simulation of basic genetic processes and systems*. The authors are grateful to G. Chirikova for assistance in translation and N.A. Kolchanov for helpful discussions.

MATHEMATICAL MODEL OF CHOLESTEROL BIOSYNTHESIS REGULATION IN THE CELL

Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A.

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

Keywords: gene network, cholesterol, regulation, mathematical model, computer analysis

Resume

Motivation:

An adequate mathematical model of the complex nonlinear gene network regulating cholesterol synthesis in the cell is necessary for investigating its possible function modes and determining optimal strategies of its correction, therapeutic included.

Results:

Dynamic model of function of the gene network regulating cholesterol synthesis in the cell is constructed. The model is described in terms of elementary processes—biochemical reactions. The optimal set of parameters of the model is determined. Patterns of the system behavior under different conditions are simulated numerically.

Introduction

Cholesterol, an amphipathic lipid, is an essential structural component of cell membranes and outer lipoprotein layer of blood serum. In addition, cholesterol is a precursor of several other steroids, namely, corticosteroids, sex hormones, bile acids, and vitamin D. Cholesterol is synthesized in many tissues from acetyl-CoA and its main fraction in blood serum resides with low-density lipoproteins (LDL). Free cholesterol is removed from the tissues with involvement of high-density lipoproteins (HDL) and transported to the liver to be transformed into bile acids. Its major pathological role is in serving as a factor causing atherosclerosis of vital cerebral arteries, heart muscle, and other organs. Typical of coronary atherosclerosis is a high ratio of LDL to HDL cholesterol [Marry R. et al., 1993]. Haploid and diploid versions of the dynamic model of function of the gene network regulating cholesterol synthesis in the cell are constructed in the work. The models are described in terms of elementary processes—biochemical reactions. The optimal set of parameters of the model allowing the calculations to comply with the published experimental data is determined through numerical experiments. Patterns of the system dynamic behavior under different conditions are simulated numerically. The results obtained are compared with the available experimental data.

Cholesterol biosynthesis and its regulation

Approximately half of the cholesterol amount present in the organism is formed through biosynthesis (about 500 mg/day) [Marry R. et al., 1993], while the other half is consumed with food. The main part of cholesterol is synthesized in the liver (~80% of the total cholesterol produced), intestines (~10%), and skin (~5%) [Klimov & Nikul'cheva, 1999].

Acetyl-CoA is the source of all the carbon atoms composing the cholesterol molecule. The main stages of cholesterol biosynthesis are described in the GeneNet database (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/).

Cholesterol regulates its own synthesis and the synthesis of LDL receptors at the level of transcription through a negative feedback mechanism [Wang et al., 1994]. A decrease in the cell cholesterol content stimulates SRP (sterol regulated protease)–catalyzed proteolysis of the *N*-terminal fragment of SREBP (sterol regulatory element–binding protein), bound to the endoplasmic reticulum (ER) membrane. On leaving the ER membrane, SREBP migrates to the cell nucleus to bind the so-called sterol regulatory element (SRE), residing in the promoter of the receptor gene, thereby switching on the receptor synthesis. In addition, SREBP activates the gene of synthase of hydroxymethyl glutaryl (HMG)-CoA reductase [Klimov & Nikul'cheva, 1999] as well as farnesyl diphosphate synthase and squalene synthase syntheses. Several studies have demonstrated rather fast effect of cholesterol on the reductase activity, unexplainable by the mere effect on the rate of enzyme synthesis. HMG-CoA reductase may be either active or inactive. Phosphorylation–dephosphorylation reactions provide for the transitions from one state into the other [Marry R. et al., 1993].

The main factors affecting the cholesterol balance at the cell level [Marry R. et al., 1993] are shown in Fig. 1.

Cell cholesterol content increases if (1) specific LDL receptors bind cholesterol-containing lipoproteins; (2) cholesterol-containing lipoproteins are bound without receptors; (3) free cholesterol, contained in cholesterol-rich lipoproteins is bound by cell membranes; (4) cholesterol is synthesized; and (5) cholesterol ester hydrolase–catalyzed hydrolysis of cholesterol esters takes place.

Cell cholesterol content decreases if (1) cholesterol passes from membranes into cholesterol-poor lipoproteins, in particular LDL_3 or LDL synthesized *de novo* (lecithin:cholesterol acyltransferase promotes this transition); (2) ACAT-catalyzed cholesterol esterification takes place; and (3) cholesterol is used for synthesizing other steroids, in particular, hormones or bile acids in the liver [Marry R. et al., 1993].

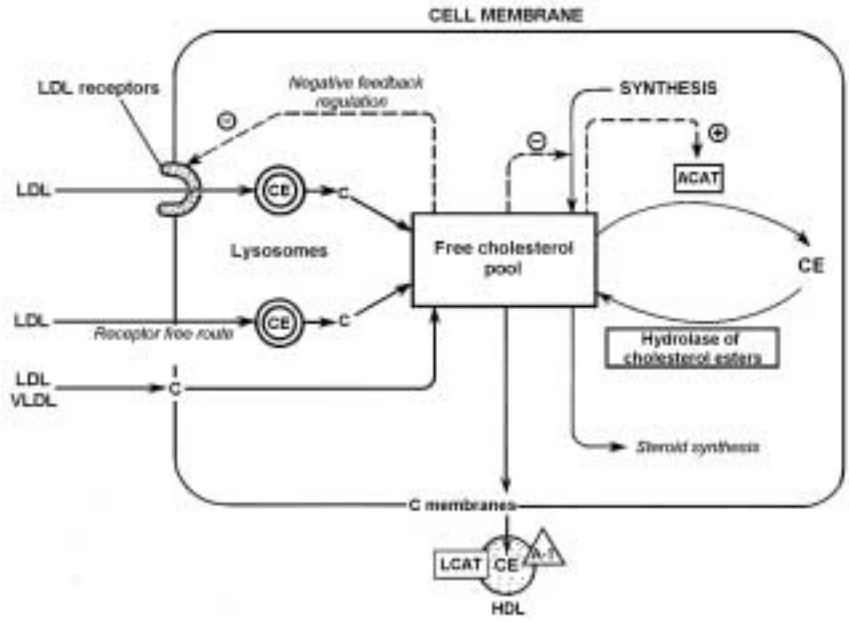


Figure 1. Factors affecting the cholesterol balance at the cell level: C, cholesterol; CE, cholesterol esters; ACAT, acyl-CoA:cholesterol acyltransferase; LCAT, lecithin:cholesterol acyltransferase; A1, apoprotein A1; LDL, low density lipoproteins; VLDL, very low density lipoproteins, HDL, high density lipoproteins; (–), inhibition of cholesterol synthesis; and (+) ACAT activation [Marry R. et al., 1993].

Methods and algorithms

A generalized chemical kinetic approach [Bazhan et al., 1995] was used for the simulation. A blockwise formalization was used, that is, each process is separated in an individual block and described independently of the other processes. A block is a simulation quantum, and its formal structure is completely described with the following three vector components: (1) X, the list of dynamic variables; (2) P, the list of constants; and (3) F, type of the right part of the system dX/dt = F(X, P) determining the rule these dynamic variables change with time. Four types of blocks are used to describe the processes in the model, namely:

Scheme 1. *Bimolecular irreversible reaction* :

$$x_1 + x_2 \underset{k_2}{\overset{k_1}{\Longleftrightarrow}} x_3 \qquad \overline{X} = (x_1, x_2, x_3), \quad \overline{P} = (k_1, k_2), \quad \frac{dx_1}{dt} = k_2 \cdot x_3 - k_1 \cdot x_1 \cdot x_2, \quad \frac{dx_1}{dt} = \frac{dx_2}{dt} = -\frac{dx_3}{dt}$$

Scheme 2. *Monomolecular irreversible reaction* :

$$x \xrightarrow{k} y_1 + y_2 + \ldots + y_n \qquad \overline{X} = (y_1, y_2, \ldots, y_n), \ \overline{P} = k, \ \frac{dx}{dt} = -\frac{dy_1}{dt} = -\frac{dy_2}{dt} = \ldots = -\frac{dy_n}{dt} = -k \cdot x.$$

Scheme 3. *Constitutive synthesis* :

$$\xrightarrow{k} x \qquad \overline{X} = x, \ \overline{P} = k, \quad \frac{dx}{dt} = k.$$

Scheme 4. *Formation of n products from m simultaneously reacting substrates* :

$\overline{X} = (x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n), \ m \geq 2, n > 0, \qquad \overline{P} = k_i, k_d,$

where k_i *is the constant of monomolecular degradation rate of the intermediate complex*,

k_d *is the Michaelis – Menten constant*.

$$\frac{dx_j}{dt} = -k_i \cdot Z, \ j = 1, \ldots, m, \qquad \frac{dy_l}{dt} = k_i \cdot Z, \ l = 1, \ldots, n, \text{ где } Z = \frac{x_1 \cdot \ldots \cdot x_m}{(k_d + x_1) \cdot \ldots \cdot (k_d + x_m) - x_1 \cdot \ldots \cdot x_m}.$$

Successive application of the blockwise approach to description of biological systems is based on the law of summation of the rates of elementary processes while uniting them in a general scheme of the simulated object development. The method of Gear [Gear, 1971] was used for numerical integration of the set of differential equations.

Results

Mathematical model

The mathematical model of intracellular cholesterol biosynthesis regulation comprises 65 kinetic blocks, 40 dynamic variables, and 93 reaction constants. The diploid model comprises 72 kinetic blocks, 44 dynamic variables, and 130 reaction constants. Experimental data, partially listed in table below, were used for the initial evaluation of certain parameters of enzymatic reactions with the system.

Table. Some constants of enzyme reactions

Enzyme	Substrate	Organism	Organ	K_c, sec^{-1}	K_m, mM
HMG-CoA reductase	HMG-CoA	Rattus norvegicus [Gil et al., 1981]	Liver	980	(-)
HMG-CoA reductase	HMG-CoA	Rattus norvegicus [Don & Kleinsek, 1979]	Liver	(-)	0.0169
HMG-CoA reductase	HMG-CoA	Rattus norvegicus [Sugano et al., 1978]	Intestine	(-)	0.0417
HMG-CoA synthase	Acetyl-CoA Acetoacetyl-CoA	Gallus gallus (hen) [Reed et al., 1975]	Liver	(-) (-)	0.1÷0.7 <0.005
HMG-CoA synthase	Acetyl-CoA	Homo sapiens [Rokosz et al., 1994]	Adrenal	(-)	0.029
Acetoacetyl-CoA thiolase	Acetoacetyl-CoA CoA	Bos taurus (calf) [Huth et al., 1975]	Liver	(-) (-)	0.01 0.025
Acetoacetyl-CoA thiolase	Acetoacetyl-CoA CoA	Gram-negative bacteria [Kim & Copeland 1997]		2.38e+4 2.38e+4	0.042 0.056
Presqualene synthase	Farnesyl diphosphate	Saccharomyces cerevisiae (yeast) [Sasiak & Rilling, 1988]		(-)	0.03
Geranyltransferase	Geranyl PP Isopentyl PP	Homo sapiens [Barnard & Popják 1981]	Liver	40.7 40.7	4.4e-4 9.4e-4
Lanosterol synthase	(R,S)-squalene-2,3-oxide	Saccharomyces cerevisiae [Balliano et al., 1992]		(-)	0.035
ACAT-1	Oleoyl-CoA Cholesterol	Homo sapiens (Cricetulus griseus)[Chang et al., 1998]	Ovary	(-)	7.4e-3
Bile acid hydrolase	Taurocholate	Lactobacillus sp. (bacteria) [Lundeen & Savage 1990]		1900	0.76

Other published data were used for evaluating parameters of the model, in particular [Klimov & Nikul'cheva, 1999]:

1. Fasting LDL concentration in adult human blood serum C_{LDL} = 200–300 mg/dl.
2. The average number of unesterified and esterified cholesterol molecules per one LDL particle Q_{UEC} = 475 and Q_{EC} = 1310.
3. LDL half-life in blood of healthy humans $\tau_{1/2}$ = 2.5 days; therefore, $k_{LDLutil.} = \ln(2)/\tau_{1/2} = 3.21*10^{-6}$ sec^{-1}.
4. Total number of LDL receptors per one cell at 37°C Q_{LDLR} = 15,000–70,000.
5. Lifespan of LDL receptors τ = 1–2 days; therefore, $k_{LDLRutil.} = 1/\tau \sim 7.72e-6$ sec^{-1}.
6. LDL receptor recyclization span $\tau \sim 20$ min.

The values of the rest parameters of the model were determined through numerical experiments.

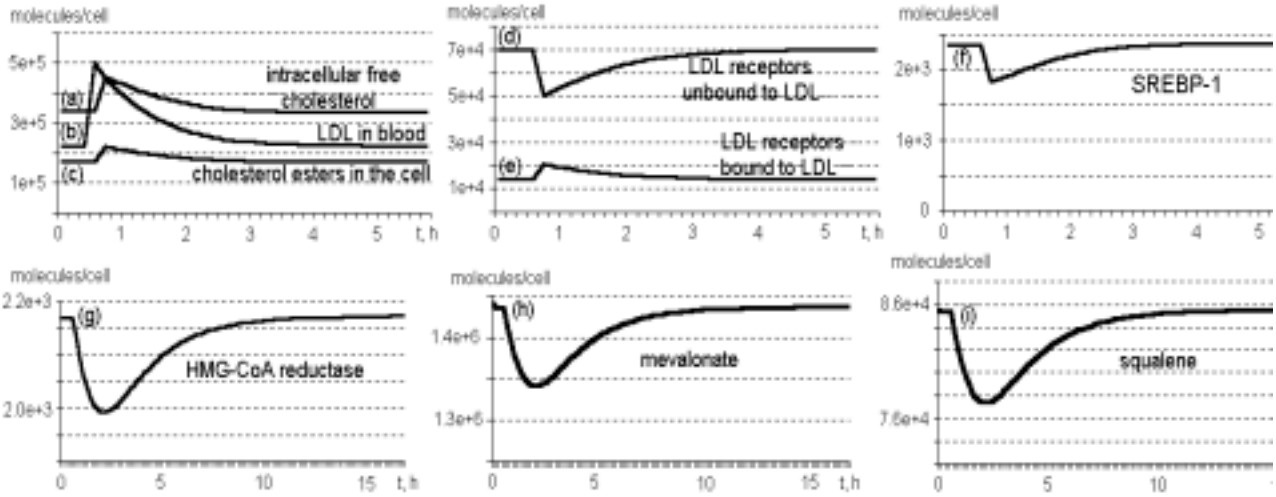


Figure 2. Kinetics of main components of the system regulating cholesterol biosynthesis in the cell.

Results of calculations

The results obtained while simulating the cell response to a twofold increase in LDL particle content in blood serum (Fig. 2, b) illustrate the model performance. The number of receptors bound to LDL increases (d); unbound, decreases (e). Intracellular concentrations of free cholesterol (a) and its esters (c) increase. Free cholesterol binds the protease (SRP), preventing SREBP-1 formation (f). Productions of enzymes involved in the internal cellular cholesterol synthesis (HMG-CoA reductase; g), LDL receptors, and intermediate low-molecular-weight components (mevalonic acid, h; squalene, i) are stopped. Cholesterol concentration in the cell is decreasing. No further influence on the system provided, it returns to the initial state. A complete recovering requires about 15 h.

In future, we plan to perform computer stimulation of recombination process in diploid cell, by modelling interactions between alleles of genes responsible for cholesterol biosynthesis.

Acknowledgments

The authors are grateful to Galina Chirikova for translation of the manuscript into English and to N.A. Kolchanov for fruitful discussions. The work was supported by National Russian Program "Human Genome" (No 106), Integrational Science Project of SB RAS "Modelling of basic genetical processes and systems".

References

1. R. Marry, D. Grenner, P. Meies, V. Roduell, "Human Biochemistry", Moscow, "Mir", (1993).
2. A.N. Klimov and N.G. Nikul'cheva, "Lipid and Lipoprotein Metabolism and Its Disturbances" St. Petersburg: Piter Kom. (1999).
3. X. Wang, R. Seto, M. S. Brown et al., "SREBP-1, a membrane-bound transcription factor released by sterol regulated proteolises" Cell, **77**, 53 (1994).
4. S.I. Bazhan, V.A. Likhoshvai and O.E. Belova, "Theoretical Analysis of the Regulation of Interferon Expression during Priming and Blocking" J. Theor. Biol., **175**, 149 (1995).
5. C. W. Gear, "The automatic integration of ordinary differential equations", Communs ACM, **14**, 176 (1971).
6. G. Gil, M. Sitges, and F.G. Hegardt, "Purification and properties of rat liver hydrohymethylglutaryl coenzyme A reductase phosphatases" Biochim. Biophys. Acta, **663,** No. 1, 211 (1981).
7. A. Don and J. W. Kleinsek, "An alternate method of purification and properties of rat liver 3-hydroxy-3-methylglutaryl coenzyme A reductase" J. Biol. Chem., **254,** No. 16, 7591 (1979).
8. M. Sugano, H. Okamatsu, and T. Ide, "Properties of 3-hydroxy-3-methylglutaryl-coenzyme A reductase in villous and crypt cells of the rat small intestine" Agr. Biol. Chem., **42,** No. 11, 2009 (1978).
9. W.D. Reed, K.D. Clinkenbeard, and M.D. Lane, "Molecular and catalytic properties of mitochondrial (ketogenic) 3-hydroxy-3-methylglutaryl coenzyme A synthase of liver" J. Biol. Chem., **250,** No. 8, 3117 (1975).
10. L.L. Rokosz, D.A. Boulton, E.A. Butkiewicz, G. Sanyal, M.A. Cueto, P.A. Lachance, and J.D. Hermes, "Human cytoplasmic 3-hydroxy-3-methylglutaryl coenzyme A synthase: expression, purification, and characterization of recombinant wild-type and Cys129 mutant enzymes" Arch. Biochem. Biophys., **312,** No. 1, 1 (1994).
11. W. Huth, R. Jonas, I. Wunderlich, and W. Seubert, "On the mechanism of ketogenesis and its control. Purification, kinetic mechanism and regulation of different forms of mitochondrial acetoacetyl-CoA thiolases from ox liver" Eur. J. Biochem., **59,** No. 2, 475 (1975).
12. S.A. Kim and L. Copeland, "Acetyl coenzyme A acetyltransferase of Rhizobium sp. (Cicer) strain CC 1192" Appl. Environ. Microbiol., **63,** No. 9, 3432 (1997).
13. K. Sasiak and H.C. Rilling "Purification to homogeneity and some properties of squalene synthetase" Arch. Biochem. Biophys., **260,** No. 2, 622 (1988).
14. G.F. Barnard and G. Popják, "Human liver prenyltransferase and its characterization" Biochim. Biophys. Acta, **661,** No. 1, 87 (1981).
15. G. Balliano, F. Viola, M. Ceruti, L. Cattel, "Characterization and partial purification of squalene-2,3-oxide cyclase from Saccharomyces cerevisiae" Arch. Biochem. Biophys., **293,** No. 1, 122 (1992).
16. C.C.Y. Chang, C.-Y.G. Lee, E.T. Chang, C.J. C.ruz, M.C. Levesque, T.-Y. Chang "Recombinant acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) purified to essential homogeneity utilizes cholesterol in mixed micelles or in vesicles in a highly cooperative manner" Journal of Biological Chemistry, **273**(52), 35132 1998.
17. S.G. Lundeen and D.C. Savage, "Characterization and purification of bile salt hydrolase from Lactobacillus sp. strain 100-100" J. Bacteriol., **172,** No. 8, 4171 (1990).

MATHEMATICAL MODEL OF ERYTHROID CELL DIFFERENTIATION REGULATION

Ratushny A.V., Podkolodnaya O.A., Ananko E.A., Likhoshvai V.A.

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

Keywords: erythroid cell, differentiation, gene network, regulation, mathematical model, computer analysis

Resume

Motivation:

Construction of an adequate mathematical model is required for investigating possible function modes of the complex nonlinear erythroid cell differentiation and maturation gene network and determining optimal strategies of its regulation while solving particular problems, therapeutic application included.

Results:

A dynamic model of the function of gene network regulating erythroid cell differentiation is constructed. The model is described in terms of elementary processes—biochemical reactions. The optimal set of the model parameter values is determined. Dynamic behavior patterns of the system under different conditions are simulated numerically.

Introduction

The hematopoietic tissue belongs to self-regenerating systems of the organism, regulated and self-regulated by specific patterns. Maintenance of a definite volume of blood cells is one of the necessary conditions of the organism function. From this standpoint, the insight into cell proliferation and differentiation of the hematopoietic tissue provided by a theoretical study is of both basic and biomedical importance. A dynamic model of the function of gene network regulating erythroid cell differentiation is constructed in the work. The model is described in terms of elementary processes—biochemical reactions. Numerical experiments are used to determine the set of parameters allowing the calculations to comply with the published experimental data. Dynamic behavior patterns of the system under different conditions are simulated numerically. The results obtained are compared with the relevant experimental data.

Regulation of erythroid cell differentiation

The main stages of erythroid cell differentiation are qualitatively represented in the GeneNet database (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/). The hormone erythropoietin interacts with the receptors of immature erythroid cells (erythroid stem progenitors of CFUe type) and stimulates their proliferation, hemoglobin synthesis, and synthesis of the enzymes involved in heme biosynthesis, that is, maturation and differentiation of erythroid progenitors [Podkolodnaya et al., 2000]. A low partial pressure of venous oxygen (hypoxia) stimulates erythropoietin synthesis.

The system regulating the erythroid cell differentiation displays a pronounced positive feedback. Erythropoietin interaction with the cellular receptor activates GATA-1 transcription factor, a key regulator of erythrocyte differentiation. GATA-1 stimulates the syntheses of α and β globins as well as the heme-synthesizing enzymes. In addition, GATA-1 activates its own gene and the gene coding for erythropoietin receptor (positive feedback). Hem, α, and β globins form hemoglobin, the major component of the mature erythrocyte.

Fe^{3+} ions are necessary for heme synthesis and, respectively, hemoglobin synthesis. Iron regulating protein (IRP) plays an important role in heme biosynthesis [Kuhn, 1994]. The system regulating influx of Fe^{3+} ions into the erythroid cell is shown in Fig. 1. It is evident from the scheme that it has not only positive, but also negative feedbacks. The specific blood serum protein transferrin transports iron from gastric mucosa and spleen sinus parenchyma into bone marrow [Fedorov, 1976]. Transferrin interacts with the membrane receptors and enters the cell through endocytosis. In the absence of Fe^{3+} ions, IRP binds the mRNA of a heme-synthesizing enzyme, 5-aminolevulinate synthetase (eALAS), reducing the synthesis. eALAS is completely unbound when the Fe^{3+} concentration in cell is high. Decrease in the globin syntheses due to certain reasons, resulting in an excess heme concentration, may inhibit the influx of Fe^{3+} either directly or indirectly through decrease in the synthesis of transferrin receptors (negative feedback) [Kuhn, 1994]. An excess of heme also has a positive effect on globin syntheses [Ponka, 1997].

Methods and algorithms

The system was described and analyzed by methods detailed in [Ratushny et al., 2000].

Mathematical model

The system regulating the erythroid cell differentiation is described with 119 kinetic blocks. The model contains 68 dynamic variables and 178 reaction constants. Values of the parameters were determined in different ways. Experimental data, partially listed in table below, were used to determine a number of enzymatic kinetic parameters of this system.

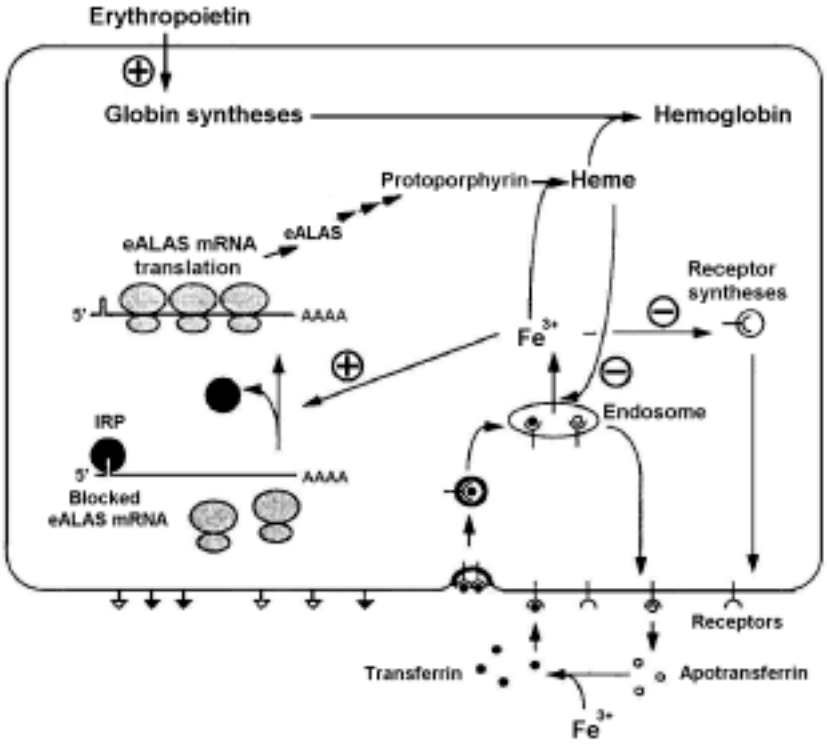


Figure 1. Regulation of Fe^{3+} influx to the erythroid cell and the control function of iron regulating protein (IRP): eALAS, erythroid 5-aminolevulinate synthetase [Kuhn, 1994].

Table. Some reaction constants.

Enzyme	Substrate	Organism	Organ	K_c, sec^{-1}	K_m, mM
Aminolevulinate synthase	Glycine Succinyl-CoA	Rattus norvegicus [Scholnick, 1972]	Liver	(-) (-)	11.0 0.07
Aminolevulinate synthase	Glycine Succinyl -CoA	Mus musculus [Tan & Ferreira, 1996]	Liver	2.05e-02 2.05e-02	8.39 1.82
Aminolevulinate synthase	Glycine Succinyl -CoA	Saccharomyces cerevisiae [Vollan & Felix 1984]		7.17e+01 7.17e+01	3.0 0.002
Porphobilinogen synthase	5-aminolevulinate	Homo sapiens [Gibbs, 1985]	Erythrocyte	11.67	0.29
Porphobilinogen deaminase	Porphobilinogen	Homo sapiens [Smythe & Williams, 1988]	Erythrocyte	2.93e-2	(-)
Porphobilinogen deaminase	Porphobilinogen with URPIII synthase	Homo sapiens [Frydman & Feinstein, 1974]	Erythrocyte	(-)	0.13 0.051
Porphobilinogen deaminase	Porphobilinogen	Rattus norvegicus [Mazzetti & Tomio, 1988]	Liver	3.43e-4	0.017
Uroporphyrinogen decarboxylase	Heptacarboxyl porphyrinogen	Homo sapiens [Mukerji & Pimstone, 1987]	Erythrocyte	1.11e-3 (9.13e-4)	2.31e-3 (7.1e-4)
Uroporphyrinogen decarboxylase	Porphyrinogen I Porphyrinogen III	Saccharomyces cerevisiae [Felix & Brouillet, 1990]		2.39 0.906	1.0e-5 6.0e-6
Coproporphyrinogen oxidase	Coproporphyrinogen III	Bos taurus [Yoshinaga & Sano, 1980]	Liver	0.203 (0.178)	0.048 (0.025)
Coproporphyrinogen oxidase	Coproporphyrinogen III	Mus musculus [Kohno et al., 1996]	Liver	0.291	0.047
Protoporphyrinogen oxidase	Protoporphyrinogen IX	Mus musculus [Dailey & Karr, 1987]	Liver	2.49	5.6e-3
Protoporphyrinogen oxidase	Protoporphyrinogen IX	Homo sapiens [Dailey & Dailey, 1997]	Placenta	1.75e-1	1.7e-3
Protoporphyrinogen oxidase	Protoporphyrinogen IX	Bos taurus [Siepker, 1987]	Liver	6.51	1.66e-2
Ferrochelatase	Protoporphyrin Fe citrate	Bos taurus [Nakahashi, 1990]	Liver	7.77e-1 7.77e-1	1.27e-2 3.51e-3
Ferrochelatase	Protoporphyrin Fe citrate	Rattus norvegicus [Taketani, 1981]	Liver	0.63 0.63	2.85e-2 3.74e-2

The values of model parameters lacking in the literature were verified through numerical experiments. At this stage, the values of parameters were selected so that the integral system behavior would comply maximally with the available experimental data on the dynamic characteristics of the system's behavior, including the following:

Hemoglobin is actively accumulated starting from the stage of basophilic erythroblast. Hemoglobin and other components are usually synthesized during G_1 and the beginning of S phases of DNA synthesis [Kozinets & Goldberg, 1982]. The total (transit) time from proerythroblast to reticulocyte formation is 120 h (5 days). Two fifth of the overall cell cycle occur during G_1 and the beginning of S phases [Fedorov, 1976]. Thus, the overall syntheses of internal components in a differentiating erythroid cell require approximately 50 h.

The number of transferrin receptors on the surface of an immature erythrocyte amounts to approximately 10^4 to 10^5 per cell; they are not detected in the progenitor cell [Kuhn, 1994].

The rates of hemoglobin synthesis in the cells at the stages of proerythroblasts and basophilic erythroblasts equal approximately 0.5 pg/h, that is, about 4.7×10^6 molecules/h for an average cell with a diameter of 10 μm; at the stage of reticulocyte, the rate decreases fivefold, that is, to about 10^6 molecules/h [Kozinets & Goldberg, 1982].

Each erythrocyte contains about 280 million hemoglobin molecules [Fedorov, 1976].

Theoretical numerical calculations according to this model demonstrate the following:

Oscillating synthesis dynamics are typical of heme, α and β globins, and the components of this molecular genetic system associated with the IRP control function and involved into Fe^{3+} influx into the erythroid cell (Fig.2, b,c,d,e). This is due to the network of negative and positive feedbacks, shown in Fig. 1.

No excess of free heme is accumulated in the differentiating erythroid cell (Fig.2, f). First, it is due to control of heme biosynthesis by IRP, which binds eALAS mRNA, and eALAS is, in turn, the enzyme involved in heme synthesis at the initial stage. Second, the excess of heme has a positive effect on globin biosyntheses.

The number of transferrin receptors on the surface of an immature erythrocyte reaches approximately 10^4 per cell, whereas these receptors are initially absent in the progenitor cell (Fig.2, c).

The rate of hemoglobin synthesis several hours after the beginning of erythropoietin effect equals approximately 5×10^6 molecules/h; maximum (8×10^6) is observed at 31 hours; in ten hours after the genes are switching off, approximately 10^6 molecules/h. Note that after 25 hours, the rate of synthesis demonstrates oscillating dinamics (calculations not shown).

An erythrocyte contains about 280 million molecules of hemoglobin on cessation of its synthesis at 74 hours (Fig. 2, d).

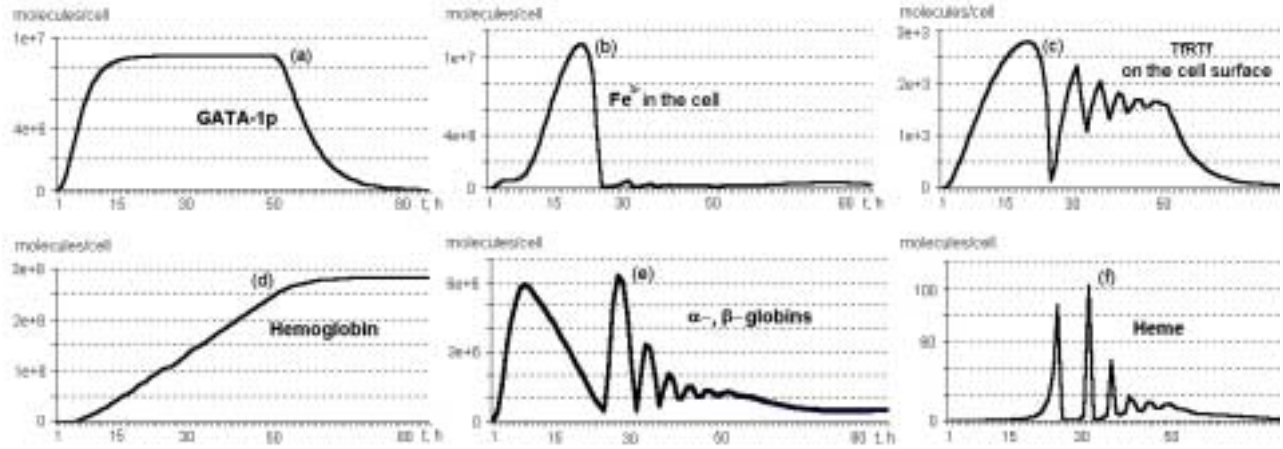


Figure 2. Dynamics of the main components on the system regulating the erythroid cell differentiation. By TfRTf are denoted transferrin receptors bound to transferrin (calculations by a model).

Acknowledgements

The authors are grateful to Galina Chirikova for translation of the manuscript into English and N.A. Kolchanov for fruitful discussions. The work was supported by National Russian Program "Human Genome" (No 106), Integrational Science Project of SB RAS "Modelling of basic genetical processes and systems".

References

1. O.A. Podkolodnaya, I.L. Stepanenko, E.A. Ananko, D.G. Vorobiev, "Representation of information on erythroid gene expression regulation in the GeneExpress system" This issue (2000)
2. L.C. Kuhn, "Molecular regulation of iron proteins." Bailliere's Clinical Haemotology, **7,** No. 4, 763 (1994).
3. Normal Hematopoiesis and Its Regulation A.N. Fedorov, Ed., Moscow, Meditsina, p. 543 (in Russian) (1976)
4. P. Ponka "Tissue-specific regulation of iron metabolism and heme synthesis: distinct control mechanism in erythroid cells" Blood, **89,** No. 1, 1 (1997).

5. A.V. Ratushny, V.A. Likhoshvai, E.V. Ignatieva, Yu.G. Matushkin, "Mathematical model of cholesterol biosynthesis regulation in the cell" This issue. (2000)
6. P. L. Scholnick, L.E.Hammaker, and H. S. Marver, "Soluble 5-aminolevulinic acid synthetase of rat liver. I. Some properties of the partially purified enzyme" J. Biol. Chem., **247,** No. 13, 4126 (1972).
7. D. Tan and G.C. Ferreira, "Active site of 5-aminolevulinate synthase resides at the subunit interface. Evidence from in vivo heterodimer formation", Biochemistry, **35,** No. 27, 8934 (1996).
8. C. Volland and F. Felix, "Isolation and properties of 5-aminolevulinate synthase from the yeast Saccharomyces cerevisiae" Eur. J. Biochem., **142,** No. 3, 551 (1984).
9. P.N.B. Gibbs, A.G. Chaudhry, and P.M. Jordan, "Purification and properties of 5-aminolevulinate hydratase from human erythrocytes" The Biochemical Journal, 1985, **230** (1), 25 (1985).
10. E. Smythe, and D.C. Williams, "A Simple rapid purification scheme for hydroxymethylbilane synthase from human erythrocytes" Biochem. J., **251,** No. 1, 237 (1988).
11. R.B. Frydman and G. Feinstein, "Studies on porphobilinogen deaminase and uroporphyrinogen III cosynthase from human erythrocytes" Biochim. Biophys. Acta, **350,** No. 2, 358 (1974).
12. M.B. Mazzetti and J.M. Tomio, "Characterization of porphobilinogen deaminase from rat liver" Biochim. Biophys. Acta, **957,** No. 1, 97 (1988).
13. S.K. Mukerji and N.R. Pimstone, "Evidence for two uroporphyrinogen decarboxylase isoenzymes in human erythrocytes" Biochem. Biophys. Res. Commun., **146,** No. 3, 1196 (1987).
14. F. Felix and N. Brouillet, "Purification and properties of uroporphyrinogen decarboxylase from Saccharomyces cerevisiae. Yeast uroporphyrinogen decarboxylase" Eur. J. Biochem., **188,** No. 2, 393 (1990).
15. T. Yoshinaga and S. Sano, "Coproporphyrinogen oxidase. Purification, properties and activation by phospholipids" J. Biol. Chem., **255,** No. 10, 4722 (1980).
16. H. Kohno, T. Furukawa, R. Tokunaga, S. Taketani, and T. Yoshinaga, "Mouse coproporphyrinogen oxidase is a copper-containing enzyme: expression in Escherichia coli and site-directed mutagenesis" Biochim. Biophys. Acta, **1292,** No. 1, 156 (1996).
17. H.A. Dailey and S.W. Karr, "Purification and characterization of murine protoporphyrinogen oxidase" Biochemistry, **26,** 2697 (1987).
18. T.A. Dailey and H. Dailey, "Expression, purification, and characterization of mammalian protoporphyrinogen oxidase" Methods Enzymol., **281**, 340 (1997).
19. L.J. Siepker, M. Ford, R. de Kock, and S. Kramer, "Purification of bovine protoporphyrinogen oxidase: immunological cross-reactivity and structural relationship to ferrochelatase" Biochim. Biophys. Acta, **913,** No. 3, 349 (1987).
20. Y. Nakahashi, S. Taketani, Y. Sameshima, and R. Tokunaga, "Characterization of ferrochelatase in kidney and erythroleukemia cells" Biochim. Biophys. Acta, **1037,** No. 3, 321 (1990).
21. S. Taketani and R. Tokunaga, "Rat liver ferrochelatase. Purification, properties and stimulation by fatty acids" J. Biol. Chem., **256,** No. 24, 12748 (1981).
22. Kinetic Aspects of Hematopoiesis G.I. Kozinets and E.D. Goldberg, Eds., Tomsk University, Tomsk, p. 306 (in Russian) (1982).

GENE NETWORK OF REDOX REGULATION AND THE PROBLEM OF INTEGRATING LOCAL GENE NETWORKS

[1]Stepananko I.L., [1]Smirnova O.G., [2]Konstantinov Yu.M.

[1]Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
[2]Siberian Institute of Plant Physiology and Biochemistry, Irkutsk, Russia
e-mail: stepan@bionet.nsc.ru
*Corresponding author

Keywords: gene networks, regulation of gene expression, transcription regulation, signal transduction pathways

Resume

Motivation:

Systematization and analysis of the miscellaneous experimental data on molecular genetic mechanisms regulating gene expression under redox changes in the cell.

Availability:

redox regulation (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet).

Introduction

Reactive oxygen species (ROS)—including superoxide radical (O_2^-), hydrogen peroxide (H_2O_2), and hydroxyl radical (–OH)—are formed in various metabolic processes. The level of ROS is an important characteristic of cell function. An excess formation of reactive oxygen species is named the oxidative stress. A variety of pathological states result from or are accompanied by an increased ROS level. Therefore, the insight into molecular mechanisms underlying the cell response to oxidative stress and maintenance of normal ROS level is of great interest.

Discovery of ROS direct involvement in intracellular signal transduction and expression regulation of the genes other than those of the antioxidant systems changed basically the conception of ROS biological role. Studies of the redox (reduction–oxidation) regulation of gene expression are becoming an actively developed direction in the field of molecular biological investigations into regulation of prokaryotic and eukaryotic genetic processes. Resulting is the concept of *redox-sensitive genes,* whose expression is efficiently regulated by the intracellular redox status.

A rapid progress in gene redox regulation studies resulted in accumulating a considerable volume of the relevant data, although yet miscellaneous, mostly in bibliographical databases. Recent studies have demonstrated the roles of redox regulation in various processes, including cell proliferation, apoptosis, and stress response. Thioredoxin and glutathione systems are involved in intracellular transducing redox signals from external factors through NADPH-dependent reduction of glutathione, thioredoxin, glutaredoxin, and redox factor to certain transcription factors, thereby casing their post-translational modification. Redox regulation may affect correct protein packing, their assembly in multimeric complexes, and binding to DNA.

The system of redox-sensitive genes is a complex ensemble of interacting genes, regulated by gene networks. The redox regulation may be considered as a natural integrator of all local gene networks, producing ROS while functioning. Therefore, the goal of this work was to describe the molecular mechanisms underlying the gene network function of eukaryotic redox regulation.

Methods

The redox regulation of gene expression is described using the computer technology GeneNet (Kolpakov *et al.,* 1998; Kolpakov and Ananko, 1999) available at http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/. The objects of gene networks and their interrelations have references to published experimental data as well as EMBL, SWISS-PROT, and MEDLINE databases. The information on transcription regulation of 60 redox-sensitive genes, their regulatory regions, and transcription factor binding sites is stored with the TRRD database (http://www.bionet.nsc.ru/trrd/), referred to in the gene descriptions.

Results

Scheme illustrating basic principles of redox regulation of gene expression is shown in Fig. 1. The scheme is visualized with the GeneNet viewer basing on the formalized information contained in the GeneNet database. This database compiles the information on signal molecules, transcription factors modified with redox changes,

and redox-sensitive genes controlling main cellular processes. The TRRD database (Kolchanov *et al.,* 2000) contains formalized descriptions of these genes.

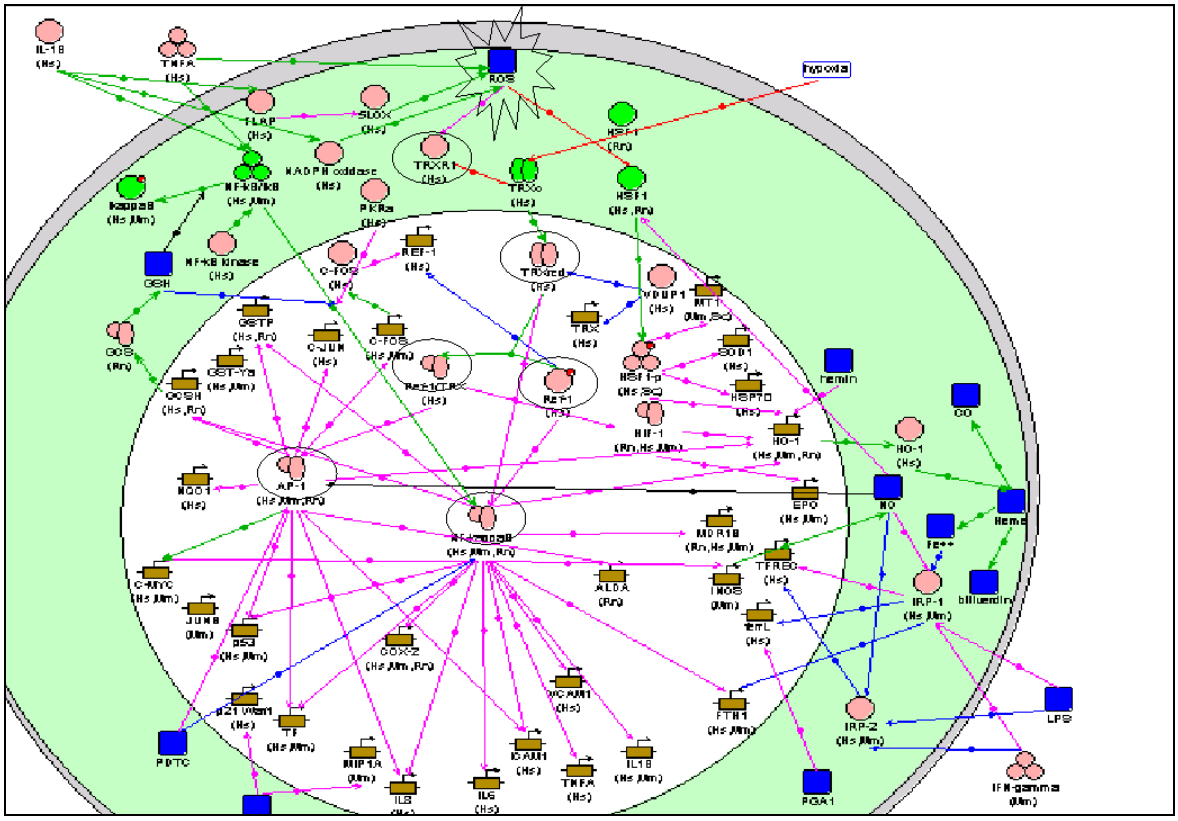


Figure 1. Scheme of redox regulation comprising the antioxidant genes induced by high ROS concentrations and the genes regulated by low ROS concentrations according to redox signaling pathway, namely, the genes involved in anti-inflammatory response.

Major activators of redox regulation gene network. Thioredoxin, one of the major factors with the thiol-mediating system, occurs in the cell in either reduced or oxidized forms and is involved in redox regulation through a reversible cysteine oxidation in its active center. In turn, the activity of thioredoxin depends on the redox status of thioredoxin reductase and its selenocysteine residues. ROS production results in selenocysteine oxidation and consequent increase in expression of this enzyme. Thus, selenocysteine is a redox sensor in the cell. Vitamin D_3–regulated protein (VDUP1), one of the known thioredoxin negative regulators, binds redox-active cysteines of the reduced thioredoxin and inhibits its biological activity as well as thioredoxin gene expression.

Another factor, the so-called redox factor Ref-1, which is a DNA repairing endonuclease (APE), is involved, together with thioredoxin, in redox regulation of DNA binding activities of several transcription factors, such as AP-1 and NF-kB. Ref-1 gene is activated by oxidative stress through induction of c-fos, the Ref-1 factor acting as a repressor of its own synthesis.

Modification of transcription factor AP-1 redox-sensitive residues regulates its DNA-binding activity and induction of a number of genes. The conservative cysteine residues in DNA-binding domains of Fos and Jun proteins mediate the redox regulation. Two proteins—thioredoxin and Ref-1, forming a heterodimer under oxidizing conditions—are necessary for AP-1 redox regulation. Thioredoxin and Ref-1 restore the NF-kB DNA-binding and transcriptional activities through interaction with cysteine at position 62 of its p50 subunit. The family of redox-regulated transcription factors includes also HSF1, a heat shock factor; p53, the product of tumor suppressor gene and key factor regulating cell cycle–related genes; and HIF-1, a hypoxia-inducible factor.

Basal response of gene network to increase in cell ROS level. ROS are formed in all cells through oxygen metabolism. Once ROS level exceeds certain threshold value, the redox regulation gene network is triggered, causing its decrease. This is provided for through activating numerous genes of the antioxidant system (genes of superoxide dismutase, catalase, glutathione reductase, peroxidases, etc.). The redox regulation is a dynamic process maintaining the balance between productions of ROS, oxidants, and antioxidants and providing for cell homeostasis. As is shown in figure, antioxidants activate the DNA-binding and transcription activities of AP-1 complex, which thereon binds to conservative ARE elements of genes.

Gene network response to increased ROS level in catabolism processes involving iron. Catabolism processes involving iron may be considered as a model of the oxidative stress. HO-1 gene expression under the effects of various stress factors is controlled by redox-regulated transcription factors AP-1, NF-kB, HIF-1, and HSF1.

Heme oxygenase is the key enzyme in heme degradation to biliverdin, carbon dioxide, and iron; in this process, biliverdin and its product bilirubin are antioxidants, whereas iron increases the oxidative stress. Ferritin binds the released iron. Regulatory proteins IRP-1 and IRP-2 control syntheses of ferritin and transferrin receptor through binding to IRE, localized to 5'- and 3'-untranslated regions of their mRNAs, according to the redox pathway via nitric oxide and cell iron content. Nitric oxide activates IRP-1 and inhibits IRP-2, an RNA-binding activity; increases ferritin expression; and decreases the level of transferrin receptor gene mRNA.

Gene network response to an ROS level increase during inflammation. NF-kB is involved in expression regulation of many genes determining the anti-inflammatory response, cell growth, and its differentiation, including those of cytokines, growth factors, and adhesins (IL-1b, IL-2, IL-6, IL-8, TNF, ICAM, iNOS, and GM-CSF). Stimulation of a cell with bacterial lipopolysaccharides, IL-1b, and TNFa results in ROS production, dissociation of inhibitor from NF-kB/IkB complex, and translocation of the active NF-kB factor into the nucleus. Depending on cell types, ROS are produced differently—through activation of either NADPH oxygenase or 5-lipoxygenase. Activation of NF-kB involves both the kinase cascade (NF-kB and IkB kinases) and redox signaling pathway, since oxidized NF-kB fails to bind to the sites of the genes it regulates. Various antioxidants, including natural antioxidant—reduced glutathione (GSH), are NF-kB inhibitors. In redox signal transduction from ROS to genes, glutathione acts as a buffer and suppresses phosphorylation of IkB inhibitor and TNFa-induced NF-kB expression. The tissue specific gene expression in response to ROS formation might be connected with different activation and binding of redox-regulated factors AP-1 and NF-kB. In case of inflammation, NF-kB activation results in induction of inducible nitric oxide synthase (iNOS) to produce nitric oxide (NO), inhibiting AP-1 DNA-binding activity.

Gene network response to an increase in ROS level under hypoxia. A low oxygen concentration (hypoxia) alters expressions of a number of genes, such as erythropoietin, heme oxygenase, and enzymes of glycolysis. The signal is transduced through thioredoxin and redox factor 1 to HIF-1 transcription factor, regulating expression of these genes. In this case, the thiol groups of cysteines are necessary for HIF-1 interaction with its coactivator CBP/p300.

Conclusion

Many key events in regulation of cellular processes, such as phosphorylation of protein transcription factors and transcription factor binding to DNA regulatory sites, are controlled by physiological redox homeostasis, in particular, thiol–disulfide balance, affected by ROS. Thus, such ROS as superoxide radical and hydrogen peroxide trigger the redox regulatory mechanism, while glutathione and thioredoxin redox systems are key expression regulators of many redox-sensitive genes, acting through changing the thiol–disulfide balance in molecules of the corresponding transcription factors. This work systematizes the available information on redox regulation of activities of p53, AP-1, NF–kappa B, HIF-1, and HSF1 transcription factors, realized through cysteine residues of DNA-binding domains with these proteins.

The redox regulation is an evolutionary conservative system controlling a vital parameter—the level of reactive oxygen species in both prokaryotic and eukaryotic cells. Any specialized gene network that appeared in the course of multicellular organism evolution had to meet the limitations imposed by the redox regulatory system. How the integration of novel gene networks into the redox regulatory system and implementation of its control function could be provided? The analysis performed demonstrates that involvement of key transcription factors controlling the function of gene networks integrated is a possible way. This integration mechanism corresponds to the principle of limiting stage in gene networks. If a gene network while functioning produces an increased ROS level, it is integrated with the redox regulatory system through its key regulatory elements—transcription factors. In this process, the following versions are possible: (1) ROS level decreasing with involvement of redox systems of major cellular biothiols—glutathione and thioredoxin; (2) the function correction (activity regulation of the system under control); and (3) cell death in case the ROS level exceeds the norm considerably.

Acknowledgements

The authors are grateful to E.A. Ananko, O.A. Podkolodnaya, E.V. Ignatieva, O.V. Kel-Margolius, and S.S. Ibragimova for annotating scientific publications in the TRRD format and to I.V. Lokhova and A.Sh. Arziev for assistance in literature search and to G. Chirikova for assistance in translation into English.

References

1. Kolchanov N.A. *et al.* (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. *Nucleic Acids Res.,* **28**, 298–301.
2. Kolpakov F.A. and Ananko E.A. (1999) Interactive data input into the GeneNet database. *Bioinformatics*, **15**, 713–714.
3. Kolpakov F.A., Ananko E.A., Kolesov G.B., and Kolchanov N.A. (1998) GeneNet: a database for gene networks and its automated visualization. *Bioinformatics*, **14**, 529–537.

PARALLEL SIMULATED ANNEALING FOR LARGE-SCALE OPTIMIZATION APPLICATIONS

Deng Y.

State University of New York, Stony Brook, USA
e-mail: deng@ams.sunysb.edu

Keywords: parallel computing, gene network

Resume

We introduce a new, efficient, and general-purpose parallelization strategy for speeding up simulated annealing on a cluster of processors connected either by dedicated switching devices or the standard Ethernet. This algorithm is applicable to broad large-scale optimization problems with continuum variables.

In this talk, we will report its effectiveness to analyzing the network of interacting genes that control embryonic development and other fundamental biological processes.

We designed a correct measure to describe and analyze parallelizing simulated annealing. We found several families of algorithms with common set algorithmic parameters that lead to optimal parallel efficiency for clusters with up to 128 processors.

Our strategy contains two major elements. First, we monitor and pool performance statistics obtained simultaneously on all processors. Second, we mix states at intervals to ensure a Boltzmann distribution of energies.

BIOLOGICAL ROLE CATEGORIES FOR REGULATORS AND MECHANISMS OF DIVERGENCE OF FUNCTION

Riley M.

Marine Biological Laboratory, Woods Hole, USA
e-mail: mriley@mbl.edu

Keywords: genomics, categories of gene products, E.coli genes

Resume

Since all of the genes of the E. coli chromosome have been sequenced, we can ask what are all the functions that are needed to code for this free-living cell. The functions of about half the genes of E. coli have been experimentally determined, and putative functions of about one-fourth of the genes have been attributed by sequence similarity to known genes.

Functions of about one-fourth of the genes remains unknown. In 1993 I arranged known E. coli genes not alphabetically but in terms of cellular function of the gene product [1]. Later, in 1996, an improved, modified scheme was included in the E.coli-Salmonella book [2]. This one-dimensional classification system has proved to be useful in the field of microbial genomics when unknown genes have been sequenced. If an unknown protein sequence showed significant similarity to one of the known E. coli gene products, function of the gene product and physiological role in the cell could be imputed.

However, we realize that the one-dimensional system does not give a complete picture of the attributes of any gene product. One can view the cellular role of each gene product from several points of view. We propose that a multidimensional system be used to categorize gene products, using six categories: metabolism, cellular processes, cell structure, regulation, transport and cell location. Any one gene can be characterized by assignments in any or all of the six major categories, and also can carry assignments to more than one slot within each major category. Some of these factors have been discussed in a minireview article [3]. Examples follow.

Degradation of lactose is a metabolic function. All genes of the lac operon including enzymes, transport and regulation are involved in that function as well as having specific roles. The enzyme beta-galactosidase is assigned to the metabolic category Degradation of Carbon compounds, specifically carbohydrates. The permease LacY belongs not only to Degradation of Carbon compounds, but also is in the Transport category, in the Major Facilitator Superfamily. The repressor LacI (or LacR) belongs not only to the metabolic Degradation of Carbon compounds but also belongs in the Regulator category in the transcriptional repressor group.

Regulation functions use different mechanisms. A chart will be shown of the different kinds of regulation. A suggested classification based on the mechanism of regulation will be presented.

Using this system, whenever function is imputed to an unknown gene by similarity to an E. coli gene of known function, a more complete view of the kind of gene and the potential physiological roles will be attached, increasing the chances of understanding the correct function of the gene in the query organism.

One can ask the question: How faithful to similarity of function are all members of a sequence-similar group of genes and proteins? Are sequence-similar groups uniform in terms of their function? Are all members of a sequence-similar group that contains a certain type of transcriptional repressor also all that type of transcriptional repressor? The example of rbsR and rbsB will be discussed, one a regulator, one a transporter.

With respect to enzymes, are all members of a sequence-similar group expected to be NADH dehydrogenases when the group contains several genes known to make NADH dehydrogenases? Many of the sequence-similar groups of E. coli proteins are all of a single function, such as acyltranferases, for instance. However, close analysis shows that some sequence-similar groups contain proteins of more than one function. Analysis of such groups may give information on mechanisms of divergence during evolution. One can imagine a group of sequence-similar genes encoding proteins of very similar function, differing only in substrate specificity, for instance. This is often the case among homologs. One can also imagine that in the course of evolution mutation can cause alteration in function such that a new capability arises. Examples of sequence-similar groups in E. coli with more than one function will be presented and discussed. One such collection is the Short Chain Reductase family.

Therefore, not only should we try to apply more than one descriptor to the function of any gene product but also existence of diversity in sequence-similar groups needs to be taken into account when assignments of function are made via sequence similarity. Being attentive to these biological matters, gradually we will be able to systematize the complexities of biology.

References

1. Riley M. (1993). Functions of the gene products of E. coli. Microbiological Reviews 57: 862-952.
2. Riley M. and Labedan B. (1996). E. coli gene products: Physiological functions and common ancestries, 2118-2202. In F. Neidhardt, R. Curtiss,III, E.C.C. Lin, J. Ingraham, K. B. Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, (ed.), Escherichia coli and Salmonella: Cellular and Molecular Biology, 2nd Ed.ASM Press, Washington, D.C.
3. Riley M. (1998). Systems for categorizing functions of gene products. Current Opinion in Structural Biology 8, 388-392.

METABOLIC ENGINEERING ELECTRONICAL INFRASTRUCTURE FOR THE DETECTION OF INBORN ERRORS

Hofestädt R.

Department of Computer Science, Otto-von-Guericke-University of Magdeburg, Magdeburg, Germany
e-mail: hofestae@iti.cs.uni-magdeburg.de

Resume

Methods of molecular biology allow the analysis and synthesis of genes, proteins, cell/protein interactions and metabolic pathways. In order to make biotechnology useful, different tools in theory and practice must be developed. This field of interdisciplinary research is called metabolic engineering. However, to solve the theoretical problems of metabolic engineering methods of Bioinformatics will help to develop and implement the electronical infrastructure. Important applications of metabolic engineering are detection of metabolic diseases, drug design and gene therapy.

Introduction

Biotechnology is important for the progress of medicine. Gene isolation and sequencing allow the detection of metabolic diseases [1]. Furthermore, proteomics support the molecular diagnosis process and methods of gene transfer are basic tools for gene therapy. Moreover, drug design can be realized using specific methods of computer science [2]. In the case of biotechnology common opinion is, that there are many problems which can only be solved using methods of different research fields: complete genetic maps are not available, the phenomena of gene regulation is not solved, the methods and tools for drug design are not available, complex information systems using molecular knowledge are not complete [3] and so forth. Fundamental is the electronical infrastructure. To understand the molecular logic of cells we must be able to analyze metabolic processes in both: qualitative and quantitative terms. In this case modeling and simulation are important and will influence the domain of medicine and (human) genetics. However, complex biochemical networks must be analyzed based on the molecular data available via the molecular database systems. The main goal of biotechnology is to modify the phenotype by recombination of the genotype using methods of molecular genetics and theoretical tools.

Metabolic Engineering

Metabolic engineering is the improvement of cellular activities by manipulation of enzymatic, transport, and regulatory functions of the cell using the recombination DNA technology [4]. The opportunity to introduce heterologous genes and regulatory elements distinguishes metabolic engineering from traditional genetic approaches to improve the strain. Metabolic engineering includes manipulation of protein processing pathways, as well as of pathways involving smaller metabolites. At present metabolic engineering is rather a collection of examples than a codified science. The main features of metabolic engineering can be separated into two parts, the theoretical part and the practical part. Synthesis of new products, creating of new products and new reactants as well as the synthesis of hybrid metabolic networks belong to the practical part. In this paper the theoretical part of metabolic engineering will be discussed.

A metabolic process within an organism is a series of enzymatic reactions consuming certain metabolites (substances) and producing others or causing changes in DNA constellation etc. The simulation of metabolic processes is based on specific models, which can be classified into the class of abstract, discrete and analytical models [5,6]. Are the abstract models based on automata and logical models which allow the global discussion of fundamental aspects, the goal of analytical models is the exact quantitative simulation, where the analysis of kinetic features of enzymes is important.

Applications

Metabolic diseases are caused by complete or partial failure of gene expression. In the domain of human genetics hundreds of genetic defects can be identified using biotechnology methods. It is more difficult to design new metabolic pathways. This task can be done using a simulator (rule based systems). The implementation of these engineering tools will be done by using the knowledge of specific database systems and information systems.

Nowadays, different subjects in the domain of medicine use methods of biotechnology for diagnosis and therapy. For example to detect bacterial or viral infections, biotechnology methods are of importance. A simple detection can be realized if the genome of this organisms is known. Therefore, hybridization using E. Coli allows the synthesis of many copies of this genetic material. Using this genome as a probe these molecules can directly be used for the identification of diseases by the hybridization of these molecules and the genome of the virus or bacteria. The result of such analysis procedures can be visualized using radioactive markers. For such hybridization experiments 10,000 to 100,000 molecules must be available. However, using methods of biocomputing only the sequences must be available. The homology test can be done automatically using methods of multiple alignment. An illness called muscular atrophy belongs to the class of metabolic diseases and appears with boys during the age of 10. The relevant gene could be identified. This gene is the biggest of all known human genes and the function of this synthetic product is the synthesis of the muscle protein Dystrophin. During the last decades it was not possible to characterize this protein. Using methods of biocomputing and protein design this protein could be characterized using the DNA sequence of this gene. This leads to the first step of causal therapy.

Electronical Infrastructure

Analysis tools need access to relevant medical and molecular data. The Metagene database [7] represents the medical knowledge of relevant inborn errors. Based on this data we developed the Metabolic Diseases Database in order to enable a simultaneous coverage of molecular and clinical knowledge about metabolic diseases [8]. A prototype of MDDB, which covered 42 diseases called hyper-ammonia (diseases with increased blood-ammonia), is available as demo for download (http://wwwiti.cs.uni-magdeburg.de/iti_bm/tools/)

in a with Borland Delphi (3.0) developed version. This version is currently translated into an Oracle database in order to provide MDDB in an Internet version (this will be ready by June/July). Knowledge is separated into a genetic and an enzymatic part. The gene-window of MDDB, covering gene locus, hereditary mode (and hence the risk of the patient's children or brothers and sisters to be affected), gene variants, gene regulation and its elements. More interesting from a clinical point of view is the information about the false enzyme. Besides general information (EC-number, name, synonyms, class membership and short description) the biochemical reactions influenced by the enzyme are listed. Substrates and products of the reaction in conjunction with structural formula and cofactors are listed too. Furthermore the compartment of the reaction is covered as well as "links" to pathways, influenced by the reaction. Those pathways, retrieved from the KEGG system [9], are illustrated using a simple graphical user interface.

Mostly neglected key factors for the success of such a system in clinics are its user interface and its handling properties. In order to get first insights, a knowledge server for the physicians at Pfeiffer Foundations in Magdeburg had been implemented. This knowledge server is managed by a server of the Institute for Technical and Business Information Systems (http://wwwiti.cs.uni-magdeburg.de/~saleski/) and for the physicians available via Internet. It offers several categories for information retrieval.

The category medical databases offers access to relevant data sets (e.g. Red Line, Medline etc.). Medical publishers and other for clinicians interesting journals are provided. Pharmanews of large pharmaceutical firms, contacts to internist Federal Associations and "links" to information about Medical IT can be found too. Another category focuses on the needs of students, physicians in education and young physicians. Because a physician should also know relevant self-help groups, a category patients' self-help is built-in as well.

The main application of the described system is the analysis of metabolic pathways. We developed a prototype, which is called Biomedical Workbench or short BioBench, which uses the concept of information fusion for the simulation of biochemical reactions [10]. In this prototype three different databases were integrated: KEGG parts of the TRANSFAC database and the described MDDB. The basis for the unique representation form is a global data model, which integrates the models of the component systems. Furthermore the kernel of the Metabolika simulation environment [11] is integrated. The user is able to specify simple queries. Hence the system integrates the query results and transfers these results automatically into the language of the simulation tool. After that the user can change the simulation parameters and start a simulation. Those results could be analyzed by a special visualization component of the prototype.

Discussion

Already the history of biotechnology shows that these new methods will directly influence different domains of medicine. However, the one important application area of biotechnology is medicine. New methods of biotechnology allow the detection of metabolic diseases, gene therapy and drug design. Therefore, biotechnology has and will change the methods of diagnosis and therapy in the area of medicine. To reach the goals of biotechnology many problems in theory and practice have to be solved. The main task belongs to the new research field of metabolic engineering, which is a subset of biotechnology [2,4,5,6]. The theoretical part of metabolic engineering opens a wide area of research problems, which can be solved using methods of Bioinformatics.

Acknowledgements

This work is supported by the german ministery of science (BMBF).

References

1. C. Scriver, A.L. Beaudet, W.S. Sly, D. Valle: The Metabolic and Molecular Bases of Inherited Disease. 7th Edition. McGraw-Hill, 1995.
2. D. Fell: Metabolic Control Analysis: a survey of its theoretical and experimental development. Biochem. J. , 286:313-330, 1992.
3. R. Hofestädt (Hrsg.): Bioinformatik 2000 - Forschungsführer Informatik in den Biowissenschaften. Berlin: BIOCOM, 1999.
4. J. Bailey: Toward a Science of Metabolic Engineering, Science, 252:1668-1674, 1991.
5. J. Collado-Vides, R. Hofestädt, M. Mavrovouniotis, G. Michal: Modeling and simulation of gene regulation and metabolic pathways. Bio Systems, Band 49, Nr. 1, S. 79-82, 1999.
6. R. Hofestädt, J. Collado-Vides, M. Mavrovouniotis: Modelling and Simulation of Metabolic Pathways, Gene Regulation and Cell Differentiation. BioEssays, 18, 1996, 333-335.
7. G. Frauendienst-Egger, F.K. Trefz: METAGENE 3.0 Computersystem zur Diagnoseunterssützung angeborener Stoffwechselerkrankungen. Wissenschaftliche Verlagsgesellschaft mbH Stuttgart, 1998.
8. R. Hofestädt, U. Mischke, M. Prüß, U. Scholz: Metabolic Drug Pointing and Information Processing. In: P. Kokol et al. (ed.), Medical Informatics Europe 99. IOS Press, Amsterdam, 1999, pp.12-15.
9. M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genoms. Nucleic Acids Research 28(1): 27-30, 2000.
10. R. Hofestädt, U. Scholz. Information Processing for the Analysis of Metabolic Pathways and Inborn Errors. BioSystems, 47, S.91-102, 1998.
11. R. Hofestädt, F. Meinecke: Interactive Modelling and Simulation of Biochemical Networks. Computers in Biology and Medicine. 25(1):321-334, 1995.

GENE CIRCUITS AND FLY SEGMENTS: SOLVING AN INVERSE PROBLEM IN DROSOPHILA

Reinitz J.B.

Mount Sinai School of Medicine, NY, USA
e-mail: reinitz@kruppel.molbio.mssm.edu

Keywords: gene expression, theoretical model, developmental genetics

Resume

Functional genomics will need to deal with all levels of biological function, including animal development. The determination of a morphogenetic field in development involves the expression of genes in spatial patterns. Sptially controlled gene expression cannot as yet be assayed in microarrays, but certain special properties of the fruit fly *Drosophila* which make it a premier system for developmental genetics also enable it to be used as a naturally grown differential display system for reverse engineering networks of genes. In this system we can approach both fundamental scientific questions about development as well as certain computational questions that arise in the analysis of genomic level gene expression data.
Our approach is called the ``gene circuit method'', and it consists of 4 components:
(1) The formulation of a **theoretical model** for gene regulation.
(2) The acquisition of **gene expression data** using fluorescently tagged antibodies.
(3) The determination of the values of parameters in the model or the demonstration that no such values exist by **numerical fits to data**.
(4) The results of (1), (2), and (3) are used (4) to **validate the model** by comparison to the existing experimental data and by making further predictions.
Recent progress in all 4 of these areas will be discussed.

BIOINFORMATIC SYSTEM IDENTIFICATION

King R.D., Garrett S.M., Coghill G.M.

Department of Computer Science, The University of Wales, Wales, UK
e-mail: rdk@aber.ac.uk
*Corresponding author

Keywords: transcriptome, proteome, metabolome, machine learning, model-based reasoning, artificial intelligence

Resume

Motivation:

There is an urgent need for new data analysis tools for transcriptome, proteome, and metabolome data.

Results:

We have developed a novel system identification tool that is capable of identifying (learning/inducing) the deep structure of a biological system from examples of its state.

Availability:

The software will be made freely available to academics after full publication in the scientific literature.

Introduction

New experimental techniques in Functional Genomics are beginning to make available large amounts of information about the transcriptome, proteome, and metabolome (TPM) of organisms. These new data are changing the focus of bioinformatics research: moving it away from the analysis of genomic sequences and towards the understanding of biological systems. New data analysis methods are needed for TPM data. To quote from DeRisi [1] on the transcriptome of *S. cerevisiae* «…the greatest challenge now is to develop efficient methods for organising, distributing, interpreting, and extracting insights from the large volume of data these experiments will provide».

The central problem in analysing TPM data is one of «system identification». Technically, system identification is the problem of selecting the model (from a set of possible models) that best fits a set of measured data points [2]. The utility of the model is measured by how well it predicts new experimental data, and by the insight that it provides. A large amount of research has been done on system identification within the field of control engineering [2]. This work is based on using quantitative time series data to construct systems of ordinary differential equations. This approach is adequate if there is a large amount of quantitative time series data available, and the set of probable forms for the system model are known. However, in cases where the data is incomplete, and the form of the model is unknown, then other methods are required. This is the case for TPM data.

Methods and algorithms

Qualitative modelling (QM) is a method of reasoning about the structure and behaviour of systems (physical or biological) which are incompletely known. It was originally devised as a means of enabling expert systems to reason from first principles. Incompleteness is dealt with by lowering the precision of the system variables to focus only on the qualitative differences in a variable's values (which in the most abstract case will be its sign). These qualitative values are formed by discretisation of the real number line (they may be symbolic, semi-quantitative or even fuzzy). QM can be seen as the first step towards developing a quantitative model. Qualitative modelling has been utilised in a number of different application domains [3], for example: diagnosis, training and control.

A number of different approaches to QM exist [3,4]. Of these the so-called «constraint based ontology» is the one which most closely resembles quantitative approaches, being an abstraction of ordinary differential equations. The qualitative simulation engine (QSIM) is the most highly developed member of this family [4]. QSIM represents systems as a set of variables and the relations between them. Each variable may take values from a quantity space consisting of a totally ordered set of landmarks and the intervals between them. QSIM is often sufficient for the synthesis of control rules [5].

TPM data comes in a form that is not suitable for traditional system identification; it is often not fully quantitative (the data often being expresses in relative units to some internal control or standard). Moreover, it has few time steps, e.g. in [1] the authors examined 6400 mRNA sequences, but only considered order of magnitude differences significant, and had only 7 time steps. Quantitative system identification methods cannot deal with

such data, qualitative ones can. Qualitative models are readily understandable by molecular biologists as the diagrams used to describe molecular systems in traditional molecular biology papers are, in effect, qualitative models. Figure 1 shows part of the RTK-Ras signalling pathway, one diagram taken from [6], a cell biology textbook, the other a qualitative model of the same pathway.

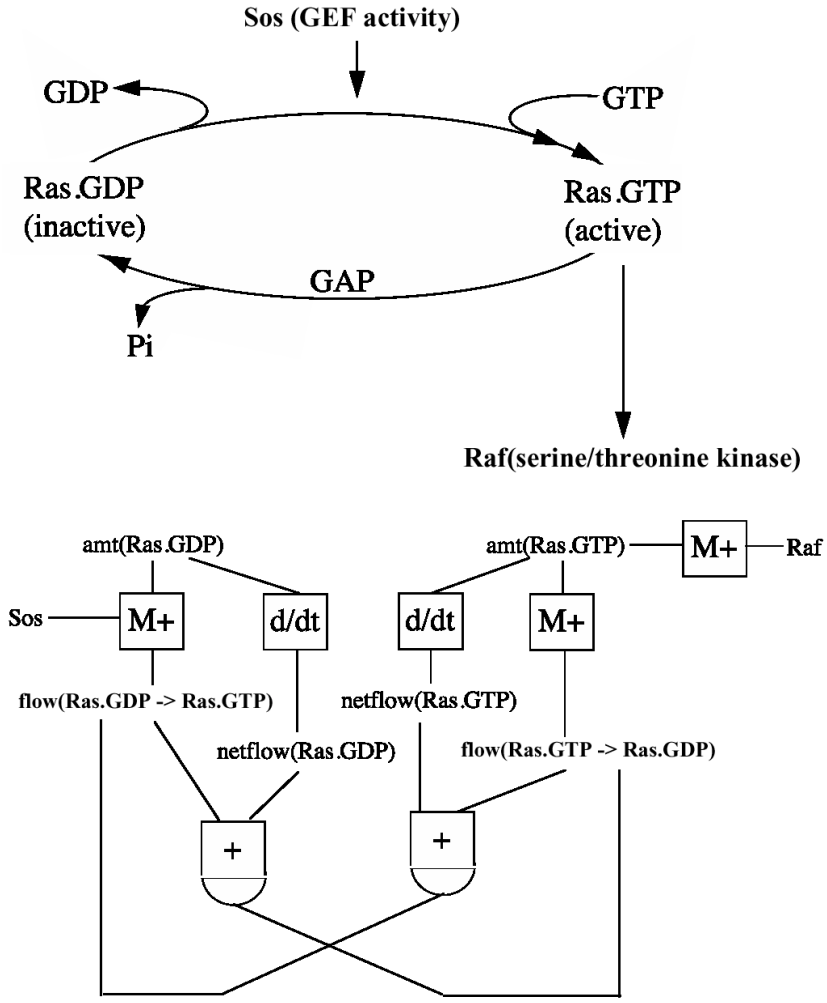


Figure 1. Inductive logic programming (ILP) [7] is the study of automated inductive learning using a first-order definite clause representation. Because its logical representation is easily translated into English, ILP systems can produce human-comprehensible output and thus are well-suited to knowledge-discovery tasks. Because first-order definite clause logic is a rich relational representation, ILP is particularly well-suited to discovery tasks where data points are best described in terms of objects (e.g. proteins, RNA molecules) and the relations that hold among them (e.g. metabolic conversions, signalling pathways). ILP has been applied to many problems in bioinformatics and drug design [8,9].

The qualitative model has the same structural form as the textbook one, but differs in being more explicit (M+ is a monotonic function, + is addition, and d/dt is derivative). Qualitative models are easier to understand than differential equation models because they are an abstraction of them. To identify a quantitative model, in typical TPM data, it is first necessary to identify a qualitative/structural one.

In recent years, the problem of identifying qualitative models has been addressed [7,18,19,20]. The process involves combining machine learning techniques with those of QM. The preferred QM approach has been QSIM because as it is constraint-based it is particularly suited to generating relations between variables, thus creating models which are abstractions of ordinary differential equations.

To learn QSIM models requires the capabilities of an ILP learning program [7]. The ILP system Progol [10] has particular advantages for identifying qualitative models. Progol performs a resource limited search through the space of possible models (that explain at least one example) relating the observed variables (and, if necessary, predicting the existence of other unmeasured variables or interactions), starting with the most abstract model and testing the predictions made by each model via QSIM against the measured data. The process terminates when the most probable model (in the Bayesian sense) that fits the data has been discovered.

Implementation and results

We have extended earlier work using Progol to identify qualitative models by:

Allowing new (hidden) variables to be introduced into models.

Use of only «positive» examples based on a Bayesian framework, i.e. use observed example behaviour only.

Incorporation of general heuristics from model theory

With this framework we are able to identify the correct systems for the standard systems: U-tube, Cascaded-tanks, Coupled-tanks, Damped Mass-Spring. We are not aware of any of any system identification problem that is as powerful.

We are currently adding TPM specific heuristics and testing the system using the gylcolysis data of Arkin et al [11].

References

1. De Risi, J.L., Vishwanath, R.I. & Brown, P.O. (1997) *Science* 278, 680-686.
2. Ljeung, L. (1987) *System Identification*, Prentice-Hall, New-Jersey.
3. Weld D, and de Kleer J. (1990) *Readings in Qualitative Reasoning about Physical Systems*, Morgan Kaufmann, Palo Alto, CA.
4. Kuipers, B. J. (1986) *Artificial Intelligence* 29, 289-338
5. Bratko, I., Muggleton, S., & Varsek, A. (1991) in *Proceeding of the Eighth International Workshop on Machine Learning*. Morgan-Kaufmann, San Mateo. 385-388.
6. Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaira, & Darnell, J. (1995) *Molecular cell biology*. Scientific American Books, Oxford.
7. Lavrac, N. & Dzeroski, S. (1994) *Inductive Logic Programming* Ellis-Horwood, Hemel Hempstead.
8. King, R.D., Muggleton, S., Lewis R.A., & Sternberg, M.J.E. (1992) *Proc. Nat. Acad. Sci. U.S.A*. 89, 11322-11326.
9. King, R.D., Muggleton, S.H., Srinivasan, A., & Sternberg, M.J.E. (1996) *Proc. Nat. Acad. Sci. U.S.A* 93, 438-442.
10. Muggleton, S. (1995) *New Gen. Comput*. 13, 245-286.
11. Arkin, A., Shen, O., & Ross, J. (1997). *Science* 277, 1275-1279.

APPLICATION OF THE METHOD OF GENERALIZED THRESHOLD MODELS FOR THE ANALYSIS OF THE EUKARYOTIC CONTROL GENE SUBNETWORKS

Tchuraev R.N., Galimzyanov A.V.

Laboratory of Mathematical and Molecular Genetics, Institute of Biology, URC RAS,Ufa, Russia
e-mail: tchuraev@anrb.ru
*Corresponding author

Keywords: eukaryotic genetic networks, control and controlled subsystems, equations, dynamics of molecular components

Resume

Motivation:

The existing analytical mathematical models of gene networks are often based on the premises unacceptable for open molecular-genetic control systems, for example, on the laws of mass action and conservation of mass. The proposed mathematical and computing means take into account the specific character of control processes at a molecular level and allow to obtain both qualitative and quantitative patterns of gene networks dynamics.

Results:

It is shown that the method of generalized threshold models enables the basic principle of functioning of eukaryotic genetic systems to be expressed correctly in terms of mathematics. The expressions describing the functioning of the controlled eukaryotic genetic block have been obtained. We have elucidated the dynamics of the control 10-gene subnetwork in the *D.melanogaster* Fly-network and control subnetwork for morphogenesis of *Arabidopsis thaliana* flower.

Introduction

Previously we proposed a method of *the generalized threshold models* (GTM) that makes it possible to plot kinetic curves for macromolecular components (DNA, RNA, proteins) in the control genetic networks of an arbitrary complexity [1]. The principal idea of the GTM method is simple enough and consists in subdivision of a molecular system of coding polymers and metabolites (м-system) into the control subsystem proper and controlled one, the former being described in terms of discrete mathematics and the latter in terms of the theory of differential equations with particular right members. The efficiency of this method was demonstrated with several prokaryotic control systems [2]. Later on a correlation was made between the formalism of the GTM method and concreteness of eukaryotic systems [3]. The present report is devoted to the research undertaken to estimate the efficiency and check the applicability of the GTM method in analyzing actual eukaryotic control gene subsystems.

Equations

1. Eukaryotic genes activity control systems have a number of peculiar features as compared with prokaryotic ones, namely, refractivity periods θ (average lifetime of the complexes “regulatory substance – DNA”) are generally much greater than for prokaryotic genomes, and characteristic times of interaction (association constants) between regulatory proteins and corresponding sites have the same order of values as for prokaryotes; in eukaryotic gene networks transcription and translation processes are spatially separated and take place in different compartments; eukaryotic mRNA are subjected to sufficiently complex process of maturing with splicing as its essential part; frequently eukaryotic genomes contain repeating sequences of genes (genes multiplicity) and have repetitive sites of specificity for regulatory molecules affecting the particular gene (sites multiplicity).

2. In the case that the m-system contains a single genome copy and the control of the j-th protein production is performed at the transcription level, we have derived the equations of dynamics in the following form:

$$\begin{cases} \dot{\mathbf{M}}_j = \mathbf{A}_{1j}\mathbf{U}_j - \mathbf{B}_{1j}\mathbf{M}_j, \\ \dot{r} = \mathbf{A}_{2j}^{T}\mathbf{M}_j - b_{2j}r_j, \end{cases} \quad (1)$$

where $\mathbf{M}_j$, $\mathbf{U}_j$ are the columns of n^j x1 dimension; $\mathbf{A}_{1j}$, $\mathbf{B}_{1j}$ are the diagonal matrices of n^j x n^j dimension; $\mathbf{A}^T_{2j}$ is the row of 1x n^j dimension; r_j, b_{2j} are the scalars, and the components $m^{(j)}_l$ of the vector-column $\mathbf{M}_j$ denote the current concentration, measured by the number of molecules per cell, of the transcripts of the 1th fraction that contain the j-th cystron; $\mathbf{U}_j = \mathbf{U}_j(t)$ is the control vector formed by the logic element and those of delay of the particular block; the diagonal element $a^{(1j)}_{ll}$ of the matrix $\mathbf{A}_{1j}$ denotes a unity intensity of the transcription beginning from one copy of the 1th promoter; accordingly, the component $a^{(2j)}_l$ of the row $\mathbf{A}^T_{2j}$ is the unity intensity of the translation from the 1th transcript; the diagonal element $b^{(1j)}_{ll}$ of the matrix $\mathbf{B}_{1j}$ is the unity intensity (degradation factor) of the 1th transcript; b_{2j} is the degradation factor of the j-th polypeptide. System (1) is a piecewise set of ordinary differential equations which is easy to be solved analytically at given values of the control vector $\mathbf{U}_j$. This set is standard for all genetic blocks in the synthesis of polypeptides controlled at the transcription level [1].

3. With the account for the peculiar features noted in Par. 1 for the simple case, when the regulatory fragment of the eukaryotic gene j consists of k_j sites of interaction with the regulatory protein of only one kind i, the equations of dynamics for controlled variables of the eukaryotic genetic block (an element of the eukaryotic control gene network) has the form:

where $m^{(0)}_j = m^{(0)}_j(t)$ is the concentration of pre-mRNA molecules containing a j-th cystron; $m^{(\vartheta)}_j = m^{(\vartheta)}_j(t)$ is the

$$\dot{m}^{(0)}_j(t) = \sum_{l=1}^{k_j} a_{lj}\, u_{lj}(t) - b_{lj}\, m^{(0)}_j(t)\ ,$$

$$\dot{m}^{(\vartheta)}_j(t) = \begin{cases} m^{(\vartheta - t)}_j(0), \text{if } t \le \vartheta, \\ m^{(0)}_j(t-\vartheta), \text{if } t > \vartheta, \end{cases} \qquad (2)$$

$$\dot{r}^{(0)}_j(t) = a_{2j}\, m^{\vartheta}_j(t) - b_{2j}\, r_j(t),$$

concentration of "mature" molecules of the j-th RNA fractions related to ribosomes; the parameter ϑ is the interval required for processing, transportation and translation of the j-th mRNA fraction. The meaning of the other denominations is the same as for set (1) [3].

4. The formalism of the GTM method has been realized in the software package [4].

Illustration of the GTM method

1. *Morphogenesis Control Subsystem (MCS) of Arabidopsis thaliana* flower. By way of example, the results of the MCS modeling for *Arabidopsis thaliana* (Fig. 1) were compared with the data gained on the basis of Mendoza and Alvarrez-Buylla model [5].

The GTM method and its program realization make it possible not only to find every possible regimes of gene network behavior, but also to test their stability under cellular divisions (Fig. 2).

2. *Subnetwork of the Drosophila Fly-gene Network*. By way of example, we have also solved the problem on the analysis, by means of the GTM method, of the early ontogenesis behavior of a fragment of the control 10-gene subnetwork in the Dr. melanogaster Fly-network that consists of the genes of three groups: gap, pair-rule and homeotic. We have elucidated the regimes of functioning the system and plotted the kinetic curves for protein gene products. When the model data is compared with the experimental results, it is apparent that the model of this subnetwork is of plausible behavior and counts in favor of the adequacy of the method applied.

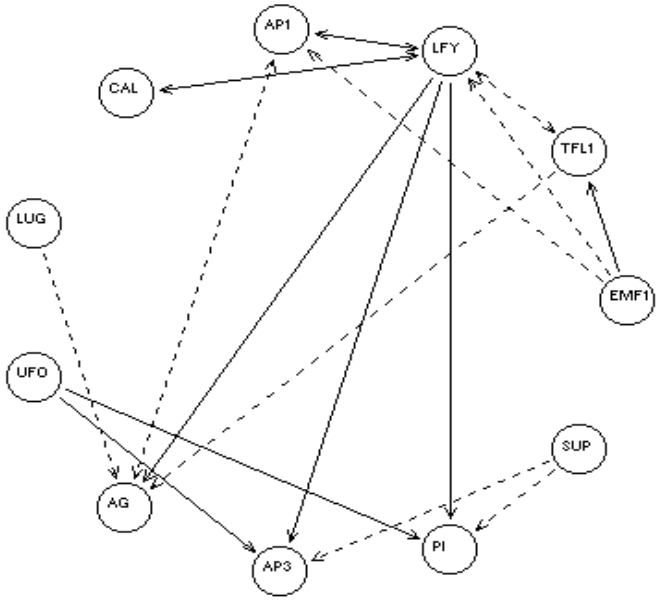


Figure 1. MCS for Arabidopsis thaliana consists of 11 genes: EMF1, TFL1, LFY, AP1, CAL, LUG, UFO, AG, AP3, PI and SUP. Generally the interactions among genes occur at the transcription level. To maintain the EMF1 gene in the active state the model simulates the effect of some external agent. BFU denotes a complex of protein products of the AP3 and PI genes, yet it is introduced into the model as an additional element [6]. The negative regulatory interactions are shown with the dotted line, and the positive one with the solid line.

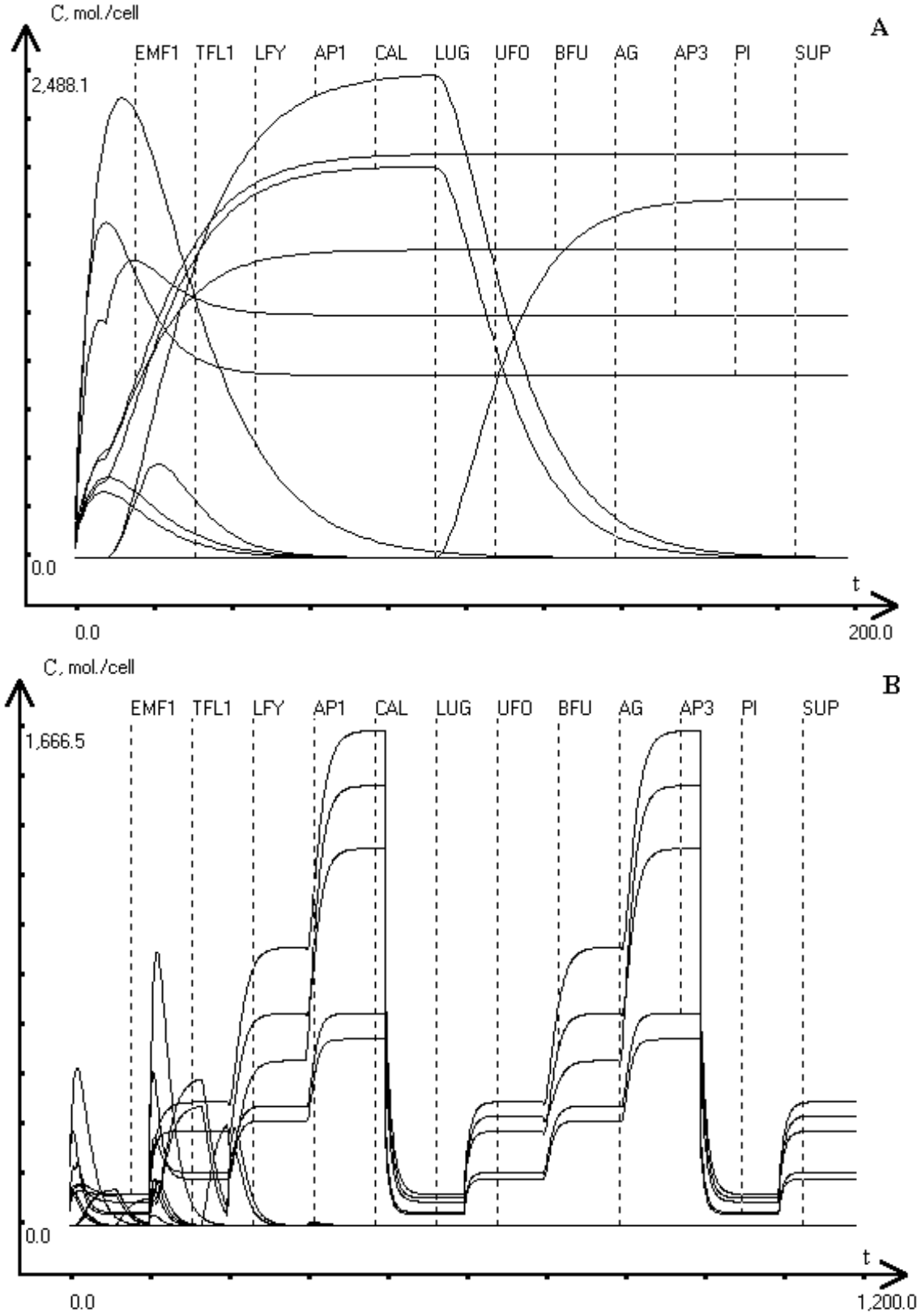


Figure 2. The dynamics of MCS functioning for *Arabidopsis thaliana* flower with no regard (A) and with regard (B) to cellular divisions and phases of a cellular cycle. At the initial point in time EMF1, AP3, PI are the switched-on genes of the system, the BFU element is active, and the other genes are switched off. After a time something like 170 steps the system turns into the stationary regime, when the EMF1, TFL1, AP3 and PI genes are switched on (concentrations of their protein products are nonzero), the BFU element is active, and the other genes are switched off. Computer experiments on the same model with regard to the phases of a cellular cycle, in the course of which the intensities of transcription and translation processes for the products of genetic blocks differ, as well as the cellular division show that there are such sets of parameters when this regime may be still retained in a series of cellular divisions. Abscissa, time; ordinate, gene product (protein) concentrations.

Conclusion

Thus, the GTM method is well suited for mathematical description of the systems of controlling eukaryotic gene expressions, since it allows to present the main regularities of functioning the eukaryotic genetic blocks in correct mathematical terms. The expressions have been derived that describe functioning of the controlled eukaryotic block. The GTM method makes it possible to reveal every possible regime of the gene network behavior, to find stationary states, to test their stability under cellular divisions, to trace the trend of variations in the number of network molecular components, to study the peculiar features of the transitions from one network functional state to another, and also to forecast the form of kinetic curves for gene products.

Acknowledgements

The research was partially supported with the RFFI grant N 98-04-49 531.

References

1. Tchuraev, R. N. (1991). A New Method for the Analysis of the Dynamics of the Molecular Genetic Control Systems. I. Description of the Method of Generalized Threshold Models. J. Theor. Biol., 151, 71-87.
2. Prokudina, E.I., Valeev R.Y., Tchuraev, R.N. (1991). A New Method for the Analysis of the Dynamics of the Molecular Genetic Control Systems. II. Application of the Method of Generalized Threshold Models in the Investigation of Concrete Genetic Systems. J. Theor. Biol., 151, 89-110.
3. Tchuraev, R.N. (1993). Method of Generalized Threshold Models for Analyzing the Dynamics of Eukaryotic Molecular-genetic Control Systems. Advanced Communication. Ufa, URC RAS, 32.
4. Galimzyanov, A.V. (2000). Software Automated Package for Analyzing the Dynamics of Control Gene Networks. In this Proceeding.
5. Mendoza, L., Alvarrez-Buylla, E.R. (1998). Dynamics of the Genetic Regulatory Network for Arabidopsis thaliana Flower Morphogenesis. J. Theor. Biol., 193, 307-319.

THE BSP-REPEATS FROM CANIDAE CONTAIN A BIDIRECTIONAL PROMOTER FOR THE RNA POLYMERASE III POTENTIALLY CAPABLE OF ENCODING DOUBLE-STRANDED RNA

*Yudin N.S., Naykova T.M., Kondrakhin Yu.V., Kobzev V.F., Romaschenko A.G.

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

Keywords: RNA polymerase III, intragenic type 2 promoter, double-stranded RNA

Resume

Motivation:

Using the standard method of contextual analysis, sequences highly homologous to the RNA polymerase III transcribed tRNA and SINE promoter elements were identified within the Bsp repeat DNAs from the canids silver fox and racoon dog. The aim was to clarify the functional activity of the computer identified potential RNA polymerase III regulatory elements.

Results:

Of the four analyzed DNA fragments, only rsV3 variant was transcribable *in vitro* by RNA polymerase III in nuclear extracts derived from HeLa cells. Assays of the transcriptional products demonstrated that a functionally active bidirectional RNA polymerase III promoter is present within rsV3 DNA. The direct and reverse strands were found to differ in transcriptional efficiency. The specific DNA structure construct is potentially capable of coding for short dsRNA performing different regulatory functions in the eukaryotic genomes.

Availability:

Detailed information about use of algorithms of contextual search for the RNA polymerase III transcribed genes are available at request. The algorithm of context search is described by Kondrakhin Yu.V. et al., (see this volume).

Introduction

Table 1. Sequence homologous to the RNA polymerase III promoter elements identified within the the Bsp repeats.

Fragment	Orientation	A box			Spacer size, bp.	B box			Transcriptional efficiency in vitro
		Sequence	Localization	S_A		Sequence	Localization	S_B	
rsV3	Direct	**GAGGTAAGTGG**	34-43	6.41	32	**AGCTCAACCCA**	75-85	5.74	**+**
	Reverse	**TAGTCTTGTGG**	64-55	6.54	41	**GCATCGGGGCT**	14-4	5.00	**+++**
	Reverse	**TGGTGCAGATGT**	242-253	6.69	51	**AGCTCAACTCT**	183-193	6.22	-
rsV51	Direct	**GAGCTAAGTGG**	322-332	7.34	32	**AGTTCAATCCA**	364-374	7.00	-
	Direct	**CAGTGGAGTGG**	610-620	7.49	-	-	-	-	-
	Direct	**GAACTAAGTGG**	1055-1065	5.47	-	-	-	-	-
	Direct	**CAGTGGAGTGG**	1344-1354	7.49	-	-	-	-	-
rsV5	Reverse	**TAGTCTTGTGG**	19-10	7.05	-	-	-	-	-
	Direct	-	-	-	-	**AGCTCAACCCA**	31-39	5.74	-
rsN2	Direct	-	-	-	-	**CGTTCAAGTCA**	625-636	7.72	-
	Direct	-	-	-	-	**TGTTCGAAACT**	741-749	7.40	-
	Direct	**AAAGGGTGTTGG**	1130-1141	5.77	30	**AGTTCAGGCCT**	1171-1179	5.88	-

Designations: S_A и S_B are the measures of similarity of Bsp repeat DNAs to the canonical A and B boxes of RNA polymerase III transcribed gene promoters [17].

The highly repeated sequences in the genomes of different *Canidae* species, we have designated as the Bsp repeats (1), show the characteristic features of regulatory type DNA (2). The major evolutionary steps of these complex hierarchical DNA molecules have most likely long preceded speciation within the *Canidae* family (3,4). The increasing abundance of their structures containing DNA elements of the regulatory type has suggested that the Bsp repeats might have been involved in the cladogenetic reorganization of the genomic material in ancestral *Canidae* (2,5). The association of satellite-like, along with interspersed-like repeats (SINE), is a feature providing indirect support for this possibility (5,6). In fact, intragenic promoter elements, characteristic of the RNA polymerase III transcribed genes (tRNA, 5S RNA, U6 RNA genes) and SINE from the genomes of various eukaryotic species, have been identified within the Bsp-repeats by computer approaches. Thus, nucleotide sequences homologous to the A and B boxes of the type 2 promoter tRNA genes have been identified (1).

Table 1 presents fragments of the Bsp repeats from the silver fox and racoon dog genomes that are homologous to the A and B boxes (3,8). The measure of similarity of each of the Bsp fragments to the canonical site DNA from the tRNA promoter regions is given in a separate column in Table 1.

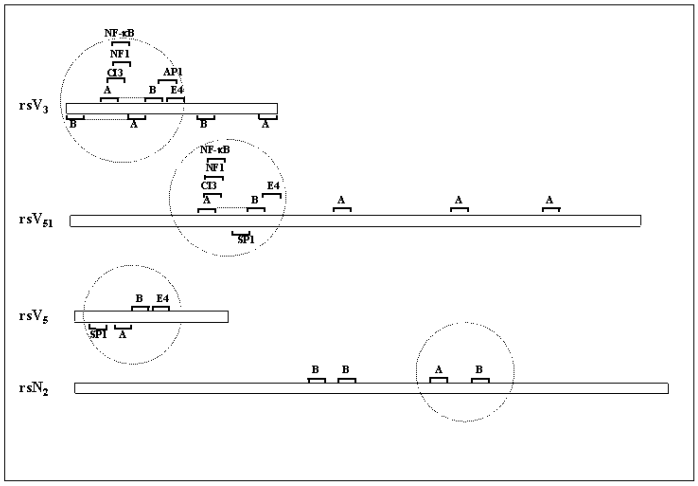


Figure 1. Scheme of position of the potential A and B boxes with respect to each other and some functional motifs capable of modulating RNA polymerase II transcription at both strands of four Bsp repeat fragments.

Fig. 1 represents schematically the position of the potential A and B boxes with respect to each other in a surrounding of functional motifs capable of, for example, modulating RNA polymerase II transcription through binding of regulatory proteins that are irrelevant to RNA polymerase III (1). The SINE-like repeat DNAs are known to associate the structural features of the RNA polymerases III and II regulatory regions, abundantly interspersed among the RNA polymerase II transcribed genes (10,11,12). Fig.1 shows the unusual combinations of regulatory RNA polymerase III elements in the Bsp repeats that make the A and B of Bsp repeats different in orientation, structure, sequence composition and spacer size from the expected canonical. This raised the question: Can the specific transcriptional factors bind to these unusual A and B boxes, and, possibly, interact to provide appropriate assembly of the RNA polymerase III transcriptional machinery and to start RNA synthesis ?

Our aim was to analyze the *in vitro* functional activity of the A and B boxes identified within the Bsp repeats by computer search.

Methods and algorithms

The potential transcriptional capacity of the Bsp repeats was tested in nuclear extracts derived from HeLa cells (13). To exclude the possible transcription of DNA sequences by RNA polymerase II, RNA was synthesized in the presence of alpha-amanitin concentration (0.5 μg/ml) that did not affect RNA polymerase III activity, yet inhibited that of RNA polymerase II (14).

Results

Using a standard system *in vitro*, in conditions optimized for RNA polymerase III transcription, DNA rsV3 from the silver fox genome within the pUV3 and pBV3 plasmids was the only one selectively transcribed. As a result, there were two transcripts of discrete size, 4.5 and 3.2 kb in pBV3. When pUV3 was used as template, the synthesized product was of different sizes. At a high concentration (200 μg/ml), alpha-amanitin inhibited the activity of both RNA polymerases III and II so that transcription of the Bsp repeats was completely suppressed. This meant that a functionally active promoter was present within the rsV3 Bsp repeat fragment.

In an attempt to determine transcript sizes, the experiments were repeated on a linear rsV3 DNA fragment. The synthesized RNA was analyzed in 1.8% agarose gel electrophoresis. As a result, transcripts of different sizes were detected. Transcripts of 60 bp were severalfold more abundant than those of 300 bp (Fig. 2).

To identify the initiating transcription strand, the synthesized products were analyzed with the primer extension assay, using reverse transcriptase. Each of the primers was complementary to one of the strands. The transcription probability for both rsV3 strands (Fig.1) could not be excluded. When oligonucleotide 5'-GCTTGTCTGAAGTGTTGTGA-3' complementary to RNA that was homologous to the reverse rsV3 strand served as primer (Fig.1), the synthesized DNA was 30 bp long. The DNAs synthesized by extension of primer 5'-GTGCATCTCTGAGGGACTTA-3' complementary to the RNA, was homologous to the direct rsV3 strand, were of different lengths 167, 150, 128, 119, 106 and shorter. This indicated that the direct and reverse strands differed in their expression levels, with one of the promoters initiating the transcription of products of different sizes.

Discussion

In vitro tests of the computer detected promoter elements within the Bsp repeat demonstrated that some of the Bsp repeat variants are transcribable by RNA polymerase III. A contextual analysis of the A and B box structure, which considered the module organization and the numerous disposition interactions within and between the modules, was performed (9,15). From the results it was concluded that Bsp repeat transcription was provided

not only by the structure of the RNA polymerase III promoter elements as such, but also by the surrounding motifs with other than RNA polymerases III regulatory functions. Thus, combinations of modules in A and B boxes (at positions 322-332 and 364-374, respectively) (Table 1) are more similar (S_A=7.34; S_B=7.00) to the canonical variants of the tRNA (9) in rsV51 as compared with rsV3 (S_A=6.41; S_B=5.74). rsV3 variant is highly homologous to rsV51 in this particular DNA region (1,3). However, rsV3 contains substitutions not occurring within the known tRNA genes in the A and B box regions. Despite this, a fully efficient transcriptional machinery is accurately assembled in this region of rsV3 DNA and not in that of rsV51 DNA. This may be accounted for by the presence of a potential site for the "ubiquitous" SP1 transcription factor between the A and B boxes in rsV51 (1). The selective SP1 binding may hinder the assembly of the active RNA transcription machinery. The absence of rsV5 transcription was expected because the boxes A and B were misoriented with respect to each other (Table 1). One plausible reason why there was no transcription in rsN2 was the strong deviation of the A and B boxes from the canonical sequences (Sa=5.77, S_B=5.88). In the A box, highly conserved adenine (A) was substituted by T at positions 7, and the slightly lesser conserved nucleotides were substituted at positions 1 (T→A) and 3 (G→A), too. Module 2 -GGG- variant (at positions 4,5,6, the A box) never occurs in combination with GTTGG sequence at the 3'-termini of the A box in certain tRNA genes.

As noted above, transcription is initiated on both RNA strands, with the promoter strength of the A and B boxes (at positions 64-55 and 14-4, respectively) of the reverse strand exceeding that of the direct strand. It is noteworthy that, of the two box pairs on the reverse strand, the one less similar to the canonical variants of the tRNA A and B boxes (at position 64-55 and 14-4) were more selective for the assembly of the initiation RNA polymerase transcription (Table 1).

It could not be ruled out that substitutions of highly conserved guanine G→T at position 12 of the A box (see Table 1, positions 242-253), as well as T→C (at positon 3 of the B box), negatively affected the functional activity of the promoter elements.

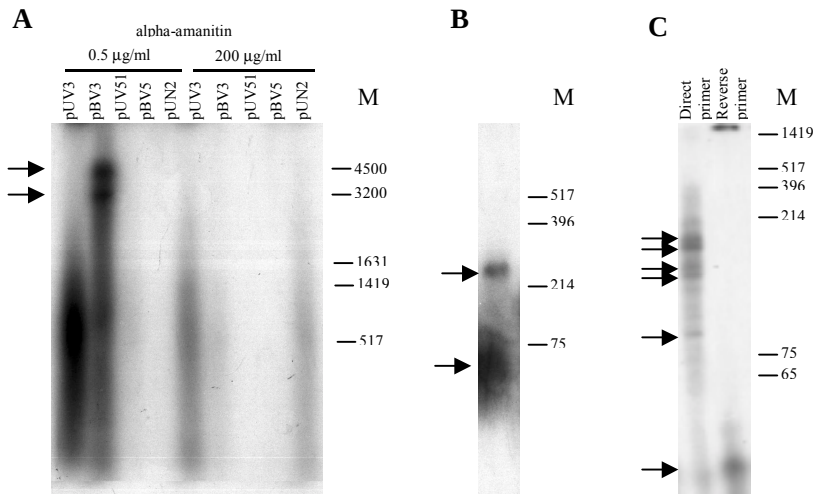


Figure 2. Selective transcription of Bsp repeats *in vitro*. (A). *In vitro* transcriptional products of different clones of Bsp repeats (pUV3, pBV3, pUV51, pUV5, pUN2) in nuclear extracts derived from HeLa cells. RNA was synthesized in the presence of the alpha-amanitin concentration (0.5 μg/ml) that did not affect RNA polymerase III activity, yet inhibited that of RNA polymerase II, and at a concentration of 200 μg/ml, that inhibited the activity of both RNA polymerases III and II. (B). Run-off transcription of rsV3 DNA fragment *in vitro*. (C). Analysis of rsV3 DNA fragment *in vitro* transcripts using primer extension.

The structural specificity of the other A and B boxes on the rsV3 reverse strand (Fig.1, positions 64-55 and 14-4 respectively) probably made this box pair the highest transcription initiating. The B box is the least similar (S_B=5.00) to the canonical of the tRNA genes with respect to structure (9). In the B box, the conserved nucleotides are substituted at positions 2 and 7 (G →C, A→G, respectively, see Table 1). The A box consists of the structural variants of modules that abundantly occur in the tRNA genes. These are module 1 (-TAG-; at positions 1,2,3, the A box), module 2 (-TCT-; at positions 4,5,6, the A box), and module 3 (-GTGG-; at positions 8,9,10,11 of the short variant of the A box) (9). The A box is unique in that -TAG- is combined with -TCT-, not with TGG-, as expected for the tRNA genes. Moreover, conserved adenine at position 7 of the A box is replaced by T in it. It is pertinent to note that the substitution of the conserved A→G at position 7 has been reported for the VAI promoter in adenovirus 2 showing the high transcriptional activity characteristic of rsV3 (16). Consequently, the distribution of specific nucleotides at each of the positions of the A and B boxes are accurately patterned to provide the appropriate transcription level.

To our knowledge, the bidirectionality of RNA synthesis for rsV3 *in vitro* (Fig.2) has not been observed for the RNA polymerase III transcribed genes (17). However, bidirectional promoters were detected in the regulatory region of the RNA polymerase II transcribed genes (18). There is a slight similarity between the rsV3 promoter regions and composed of two oppositely oriented and quite far apart tRNA genes of the trypanosome promoter (19). The DNA region identified within rsV3 with overlapping A and B box pairs on complementary strands appears to be potentially capable of coding for short double stranded RNA (dsRNA). It is known that dsRNAs regulate eukaryotic genome functions by activating specific PKR family protein kinases that inhibit translation process (20).

Screening for sequences homologous to the bidirectionally promoter construct we found within rsV3 in the GeneBank revealed no such sequences. The potentially significant regulatory role of the DNA constructs as sources of genes coding for the nontranslated dsRNAs should be kept in mind. This warrants further screening rounds of the GeneBank that would include sequence data for full size eukaryotic genomes. The absence of homology of rsV3 with the known tRNA genes, the ancestors of many SINE sequences, is evidence that the rsV3 RNA polymerase III promoter boxes might have arisen otherwise than those in SINE sequences. Taken together, the hierarchical structure of the Bsp repeats, the abundance of direct and inverted short repeats and numerous unusually combined RNA polymerase III elements (Table 1, Fig. 1) strongly suggest that a convergent mechanism underlies the evolutionary emergence of the promoter elements within the Bsp repeats.

Acknowledgments

This work was supported partly by the Russian Foundation for Basic Research grant 99-04-49722. We are grateful to Prof. K. Seifart for providing us reagents and interest in the work.

References

1. Potapov V.A., Solov'ev V.V., Romaschenko A.G., Sosnovtsev S.V., Ivanov S.V. (1991) Features of the structure and evolution of the complex tandemly organized Bsp repeats in the fox genome. II. Tissue-specific and recombination sites of the BamHI dimer. *Mol.Biol. (Msk),* 25, 116-132.
2. Romaschenko A.G., Yudin N.S. (2000) Molecular evolution of Bsp repeats from *Canidae:* structural development and mechanisms of DNA informational enrichment. *In "Modern concepts of evolutionary genetics" V.K. Shumny and A.L. Markel eds., Novosibirsk, ICG SD RAS,* 177-188 (in Russian).
3. Potapov V.A., Solov'ev V.V., Romaschenko A.G., Sosnovtsev S.V., Ivanov S.V. (1990) Structure and evolution of complex tandem Bsp repeats in fox genome. I. Structure and internal organization of a BamHI dimer*. Mol.Biol. (Msk*), 24, 1649-1665.
4. Sosnovtsev S.V., Ivanov S.V., Solov'ev V.V., Potapov V.A., Romaschenko A.G. (1993) Recombination events in evolution of satellite-like Bsp repeats: formation of subrepeat and monomer units predates the divergence of *Canidae* lineages. *Mol.Biol. (Msk),* 27, 992-1013.
5. Belyaeva T.A., Vishnivetsky P.N., Potapov V.A., Zhelezova A.I., Romaschenko A.G. (1992) Species- and tissue-specific transcription of complex, highly repeated satellite-like Bsp elements in the fox genome. *Mammalian Genome,* 3, 233-236.
6. Mikhailova S.V., Babenko V.N., Romashchenko A.G., Beliaeva T.A., Melidi N.N., Lavrinenko V.A., Guvakova T.V., Ivanova L.N. (1995) Distribution of Canidae Bsp-repeat transcripts in arctic fox kidney: structural similarity of Bsp-repeats with SI NEs. *Dokl Akad Nauk*, 343, No. 2, 260-264.
7. Potapov V.A., Sosnovtsev S.V., Solov'ev V.V., Ivanov S.V., Romaschenko A.G., Kolchanov N.A. (1988) Structure of complex repeats from fox C-heterochromatin: regulatory elements of replication, recombination and gene expression control. *Dokl Akad Nauk,* 299, No. 5, 1250-1255.
8. Ivanov S.V., Potapov V.A., Filipenko E.A., Romaschenko A.G. (1991) Heterogeneity of the *Canidae* Bsp repeat family: discovery of the EcoRI subfamily. *Genetika,* 27, No.6, 973-982.
9. Rogozin I.B., Kondrakhin Yu.V., Naykova T.A., Yudin N.S., Voevoda M.I., Romaschenko A.G. This issue..
10. Weiner A.M., Deininger P.L., Efstratiadis A. (1986) Nonviral retroposones: genes, pseudogenes and transposable elements generated by the reverse flow of genetic information. *Ann. Rev. Biochem.,* 55, 631-661.
11. Britten R.J. (1997) Mobile elements inserted in the distant past have taken an important functions. *Gene,* 205. 177-182.
12. Britten R.J. (1996) DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. USA,* 93, 9374-9377.
13. Weil A.P., Segall J., Harris B., Ng S.Y., Roeder R.G. (1979) Faithful transcription of eukaryotic genes by RNA polymerase III in systems reconstituted with purified DNA templates. *J. Biol. Chem..,* 254, No. 13, 6163-6173.
14. Transcription and translation. A practical approach. (1987) B.D. Hames and S.J. Higgins eds., IRL Press, Washington.
15. Kondrakhin Yu.V., Rogozin I.B., Romaschenko A.G. *This issue*.
16. Schneider H.R., Waldschmidt R., Seifart K.H. (1990) Human transcription factor IIIC contains a polypeptide of 55 kDa specially binding to Pol III genes. *Nucleic Acids Res.,* 18, No.16, 4743-4750.
17. Geiduschek P.E., Tocchini-Valentini G.P. (1988) Transcription by RNA polymerase III. *Ann. Rev. Biochem.,* 57, 873-914.
18. Romaschenko A.G., Potapov V.A., Solov'ev V.V. (1989) Distribution, structure and functional significance of complex repeated DNA sequences from animal genomes. *In "Structural-functional organization of genome" Ed. by V.K. Shumny, Novosibirsk, Nauka*, 80-114.
19. Nakaar V; Dare AO; Hong D; Ullu E; Tschudi C (1994) Upstream tRNA genes are essential for expression of small nuclear and cytoplasmic RNA genes in trypanosomes. *Mol Cell Bio.,* 14, No.10, 6736-6742.
20. Proud C.G. (1995) PKR: a new name and new roles. *Trends Biochem. Sci.,* 20, 241-246.
21. Smit A.F., Riggs A.D. (1995) MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. . *Nucleic Acids Res.,* 23, No. 1, 98-102.

MODELING OF CELL CYCLE GENE REGULATORY NETWORK. A ROLE OF A POSITIVE FEEDBACK LOOP IMPLYING POTENTIAL E2F TARGET SITES IN THE REGULATORY REGIONS OF AP-1 GENES

Deineko I.V., Kel-Margoulis O.V., Ratner V.A., Kel A.E.

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

Keywords: dynamic modeling, cell cycle, gene network, regulation, E2F binding sites

Resume

Motivation:

Molecular mechanisms underlying programs of cell proliferation, differentiation and apoptosis are now under intensive investigation. Recent studies strongly suggest that triggering of these programs strongly depends on the cell cycle control. Integrative approaches such as computer modeling are needed to analyze different modes of the complex dynamic of the cell-cycle gene network.

Results:

A considerable amount of data on the cell cycle gene control have been collected from the scientific literature and was compiled in the CYCLE-TRRD database (http://wwwmgs.bionet.nsc.ru/mgs/ papers/kel_ov/celcyc/). These data serve as a basis for the reconstruction of the gene regulatory network controlling G1/S transition. We have performed the dynamic modeling of this regulatory network. Two alternative modes corresponding to the S-phase entry or exit into the G0 are obtained. We have shown that the switch between these two modes depends on the duration of a mitogen stimulation. The positive feedback loop through E2F sites computationally predicted in the promoters of the AP-1 family genes found to be very important for the S-phase entry.

Introduction

Extremely complex regulatory interactions among genes providing cell cycle progression are in focus of intensive investigations. There are still a lot of “white spots” in our understanding of the control of different cellular programs: proliferation, differentiation and apoptosis. Recent studies strongly suggest that triggering of these programs strongly depends on the cell cycle control. Integrative approaches such as computer modeling are needed to develop dynamic models of the cell-cycle gene network functioning.

Transcription factors of the E2F family play an important role in the regulation of cell cycle progression. E2F/DP heterodimers control transcription of genes whose expression is maximal at the G1/S boundary and at the beginning of the S-phase. Activity of the E2F factors can be repressed by the retinoblastoma protein. In hypophosphorylated state pRB specifically bind to and masks the activation domain of the E2F-1 factor. For cell to progress through G1/S boundary pRB protein should be phosphorylated by cyclin-dependent kinases (cdks). Activity of cdks is regulated at several levels: cell cycle-specific transcriptional regulation, formation of complexes with regulatory subunits – cyclins, dephosphorylation by phosphatase Cdc25A.

A considerable amount of data on the cell cycle gene control have been collected from the scientific literature and was compiled in the CYCLE-TRRD database (http://wwwmgs.bionet.nsc.ru/mgs/ papers/kel_ov/celcyc/). These data serve as a basis for reconstruction of the cell cycle gene regulatory network. Moreover, data presented in the CYCLE-TRRD database were used to develop computer methods for E2F binding site recognition and to enrich the cell cycle gene network by potential E2F target genes. In this work we consider a basic gene regulatory network particularly controlling G1/S transition. We have performed a dynamic modeling of the network that included a minimal number of E2F regulated genes. A computer-predicted E2F target genes were introduced in the model in addition to the well characterized genes regulated by E2F family. These new E2F target genes are shown to be important for understanding the dynamic of the network in response to the proliferative signals.

Results and Discussion

Computer prediction of E2F sites in genes coding AP-1 components

We have developed a method for recognition of potential E2F sites in gene regulatory regions and for the identification of E2F target genes (Kel et al., 1999). Applying this method we have proposed a number of new potential E2F target genes (see the list of these genes in http://compel.bionet.nsc.ru/cell-cycle/). Among them there are some genes that code various components of the AP-1 transcription factors: c-fos, c-jun, fosB, fra-1, fra-2, junB, junD. AP-1 factors are known to mediate gene activation in response to wide variety of extracellular signals. Potential E2F binding sites within *jun* and *fos* genes suggest a possible existence of a positive feedback loop from E2F-1 factor to AP-1 genes.

Model of the network controlling G1/S transition

To analyze the role of the revealed E2F sites in the regulation of cell cycle we have developed a dynamic model of G1/S transition under mitogen stimulation (Fig. 1). The model includes gene regulation by E2F-1 and pRB proteins as well as biochemical reactions such as phosphorylation and dephosphorylation. We have considered positive transcriptional regulation by E2F-1 for the following genes: E2F-1 itself, tumor suppressor gene pRB, cyclin E and AP-1 components. Retinoblastoma product is considered in two functional states: active hypophosphorylated (pRB) and inactive hyperphosphorylated (pRB-p). We have included in the scheme the coupled phosphorylation-dephosphorylation (PD) cycles between cyclinE/cdk2 and Cdc25A that was proposed in the work of Aguda and Tang (1999) (transition between the active (a) and inactive (i) cyclinE/cdk2 complexes, see the Fig.1). The mitogen activation is modeled by an impulse function of the AP-1 concentration (see below). Potential regulatory interrelations between E2F-1 and AP-1 genes are taken into account as an additional positive feedback loop in the model (shown with the question mark). We have analyzed dynamic behavior of the model developed as well as an influence of the additional feedback loop.

In the Fig. 2 one can see a system of kinetic equations for the considered model. Concentrations of proteins and complexes ($y_1,\ldots,y_6$) depend on their production, modification and degradation.

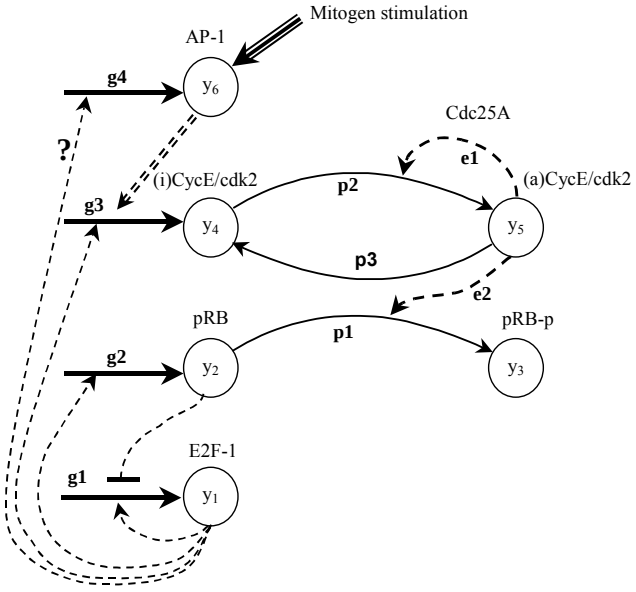


Figure 1. A model of regulatory network providing G1/S transition in cell cycle.

The thick arrows show the processes of protein production from the corresponding genes (processes no: g1,g2,g3,g4);

thin solid arrows – processes of protein phosphorylation/ dephosphorylation (processes no: p1,p2,p3);

thick dashed arrows – enzymatic catalysis (processes no: e1,e2);

thin dashed arrows – positive regulation of gene production by E2F transcription factors;

thin dashed unsharpened line – repression by pRB;

double dashed arrow – activation of Cyclin E after mitogen stimulation though AP-1 factors.

$$\frac{d}{dt}y_1 := \frac{e_{max} \cdot k_1 \cdot y_1}{k_1 \cdot y_1 + \left(k'_1 + k''_1 \cdot y_1\right) \cdot y_2} - \phi_1 \cdot y_1$$

$$\frac{d}{dt}y_2 := k_2 \cdot y_1 - k_3 \cdot y_2 \cdot y_5 - \phi_2 \cdot y_2$$

$$\frac{d}{dt}y_3 := k_3 \cdot y_2 \cdot y_5 - \phi_3 \cdot y_3$$

$$\frac{d}{dt}y_4 := k_4 \cdot y_1 + k''_4 \cdot y_6 - k_{4i} \cdot y_4 \cdot y_5 + k_{4a} \cdot y_5 - \phi_{4i} \cdot y_4$$

$$\frac{d}{dt}y_5 := k_{4i} \cdot y_4 \cdot y_5 - k_{4a} \cdot y_5 - \phi_{4a} \cdot y_5$$

$$\frac{d}{dt}y_6 := k_6 \cdot y_1 + F_6(t) - \phi_6 \cdot y_6$$

Figure 2. A system of kinetic equations for the model of G1/S transition. y1,…y6 – concentrations of proteins and complexes; k – speed constants; ϕ - degradation constants; F6 – mitogen activation function.

Comparison of dynamic behavior of the model with and without positive feedback loop to AP-1

We have performed a number of computer simulations of the system of equations applying various sets of kinetic constants and initial parameters. The impulse function $F_6(t)=Fm$ if $t<tm$ and $F_6(t)=0$ if $t\geq tm$ was applied to simulate mitogen signaling. Here, tm is the duration of cell stimulation by mitogen and Fm – its strength. First, we have analyzed the dynamic behavior of the model where possible feedback loop from E2F-1 factor to AP-1 genes is taken into account. Two alternative modes were revealed depending on the duration of mitogen stimulation (see Fig. 3). If the duration is relatively long (tm>10 units) the consequent increase of AP-1 concentration and active CycE/Cdk2 complexes is observed even after the mitogen activation is turned off. This switches the system to an autonomous mode when concentration of E2F-1 factor is increased and reaches high values required for the S-phase entry (Fig. 3a). If the duration is relatively short (tm<10) then AP-1 concentration is going down immediately after turning off the mitogen stimulation and concentration of E2F-1 is stabilized at a low levels equilibrated by the constant concentration of pBR. This mode corresponds to the cell exit into the G0 state (Fig. 3b).

Then we have analyzed the dynamic behavior of the model without consideration of the positive feedback loop from E2F-1 factor to AP-1 genes. The feedback loop was cut off by setting $k_6=0.0$. In this situation the autonomous mode of the system and S-phase entry could be observed after very long mitogen stimulation (tm>54) (Fig. 4a).

Thus, we have shown, that the additional positive feedback loop is able to modulate significantly dynamic behavior of cell cycle gene regulatory network. Our results suggest an existence of the mode when activation of the positive feedback loop from E2F-1 factor to AP-1 genes lead to the S-phase entry just after short mitogen stimulation. This mode can be used for further modeling of triggering program of uncontrolled cell proliferation and ultimately of a tumor development.

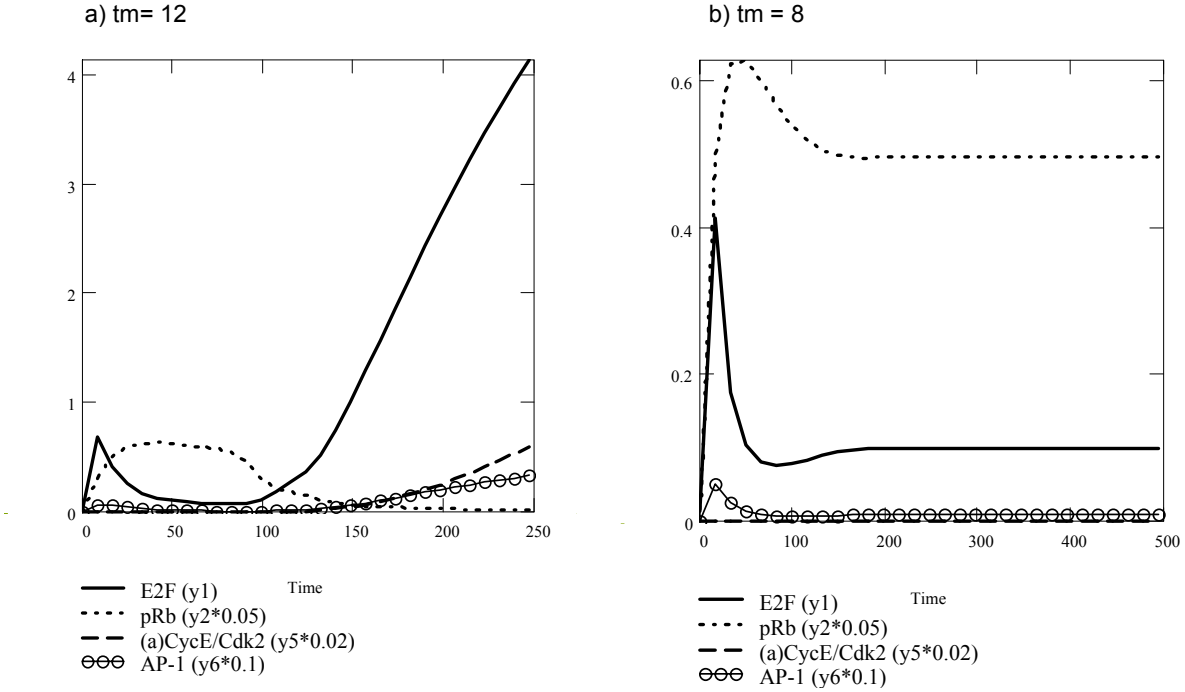


Figure 3. Two modes of the dynamic behavior of the model *with* positive feedback loop from E2F-1 factor to AP-1 genes. Rate constants are: $k_1=k'_1=k_2=k_{4i}=1$, $k''_1=10$, $k_3=0.4$, $k_4=0.09$, $k_{4a}=2$, $k''_4=0.1$,$k_6=0.085$ (*with* AP-1 feedback loop), $e_{max}=2$, $\phi_1=\phi_3=\phi_6=0.1$, $\phi_2=\phi_{4i}=\phi_{4a}=0.01$, Fm=0.044, initial values: $y_1=0.014$, $y_2=0.006$, $y_3= y_4= y_6=0$, $y_5=0.0001$, **a)** Mode corresponding to the S-phase entry, duration of the mitogen stimulation is relatively long (tm = 12); **b)** Mode corresponding to the exit into the G0 state, duration of the mitogen stimulation is relatively short (tm = 8).

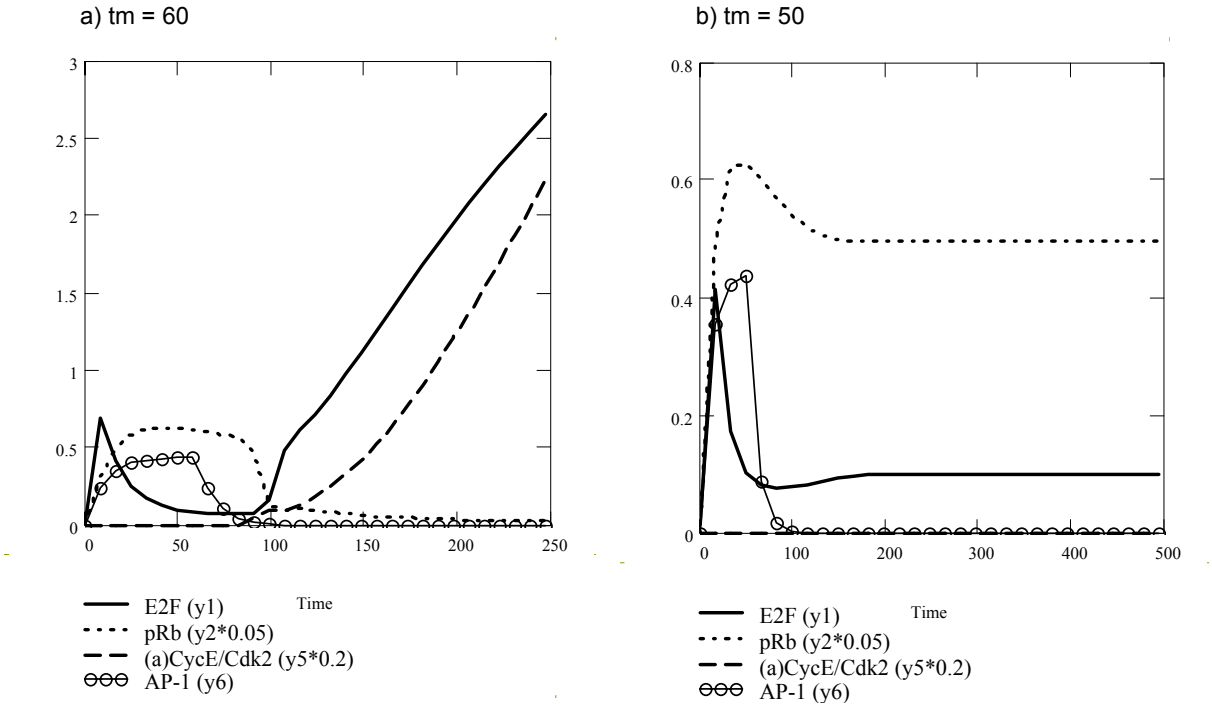


Figure 4. Two modes of the dynamic behavior of the model *without* positive feedback loop from E2F-1 factor to AP-1 genes. Constants are the same as in the Fig. 3, except k_6=*0.0.* **a)** Mode corresponding to the S-phase entry, duration of the mitogen stimulation tm = 60; **b)** Mode corresponding to the exit into the G0 state, duration of the mitogen stimulation tm = 50.

Acknowledgements

Different parts of this work were funded by the Siberian Branch of Russian Academy of Sciences and by Volkswagen-Stiftung (I/75941).

References

1. Aguda B.D., Tang Y, The kinetic origin of the restriction point in the mammalian cell cycle (1999) *Cell Prolif.*, 32, 321-335.
2. Kel A., Kel-Margoulis O, Wingender E. E2F composite units in promoters of cell cycle genes. (1999) *In Proceedings of the German Conference on Bioinformatics GCB'99, October 4-6, 1999,* Hannover, Germany, 148-154.

THE INTEGRATED TRANSFAC SYSTEM AS A BASIS FOR MODELING AND SIMULATION OF GENE REGULATION MECHANISMS

Potapov A., Christensen M., Drewes V., Schacherer F., Wingender E.

GBF German Research Centre for Biotechnology, Braunschweig, Germany
e-mail: frs@gbf.de
* Corresponding author

Keywords: regulatory networks, signal transduction, transcriptional regulation, database integration, TRANSFAC database

Resume

Motivation:

Control of gene expression appears to occur mainly at transcriptional level, but this is frequently in response to extracellular messengers. However, information required to construct comprehensive pictures of the underlying information is still fragmented.

Results:

Therefore, we decided to integrate our database resources for transcriptional regulation and signal transduction, along with a number of additional databases, to a more comprehensive resource, the TRANSFAC system. This has been done under a relational database management system that integrates now the databases TRANSFAC, TRANSPATH, CYTOMER®, S/MARt DB and PathoDB. On that basis, approaches to simulate the dynamics of gene regulation mechanisms appear to be promising.

Availability:

Parts of the TRANSFAC system are available online at (http://transfac.gbf.de). Access is free for users from non-commercial organizations.

Introduction

Regulation of gene expression is a highly complex issue, occurring at a number of distinct levels in terms of structural and functional hierarchies. Considering transcriptional control as the main level for gene regulation, provided by a huge number of transcription factors, their activities are highly dependent on the tissue, developmental stage and conditional factors under consideration. Several databases are available on the WWW that describe transcriptional regulation mechanisms from slightly distinct viewpoints: TRANSFAC (Knüppel et al., 1994; Wingender et al., 2000), TRRD (Kolchanov et al., 2000), PlantCare (Rombauts et al., 1999), PLACE (Higo et al., 1999). TRANSFAC contains textual information about expression patterns of transcription factors and about the regulation of their activities, but these data are hitherto poorly structured.

To enable proper modeling of the gene expression "coordinates" mentioned above, we have prepared a number of databases that allow modeling expression locations (organs/tissues and cell types), expression stages (developmental stages) and conditions. In addition to several genuine functions of these data resources, they represent controlled vocabularies to map expression patterns, for instance of transcription factors, and are being extracted to establish comprehensive ontologies of "expression states". For that purpose, an overall integration schema has been developed for these data resources that will provide a basis to simulate the dynamics of gene expression mechanisms.

Methods and algorithms

TRANSFAC is a relational database that is running under MS SQL-Server 6.5. The same is true for S/MARt DB that is tightly linked to TRANSFAC. TRANSPATH has been originally created as an object-oriented database. It has been established under an object-oriented database management system (from POET Software, Hamburg). As an interface to the database, Java and Object Query Language (ODMG 1997) are used in servlets, providing access over the WWW. The database has been transformed to the MS SQL-Server 6.5 as well and, therefore, a relational version of TRANSPATH is available now. CYTOMER is a relational database, it has been initially established on the miniSQL platform and then was transformed to mySQL. PathoDB was established and maintained as a stand-alone version under MS Access, and has been independently shifted to the MS SQL-Server 6.5 before its integration.

For some variants of kinetic modeling of signal transduction pathways we are using the method of directed graphs developed for the analysis of complex enzymatic systems (Volkenstein and Goldstein, 1966).

Implementation and results

TRANSFAC is a well-known and widely used database on transcription factors, their genomic binding sites and their DNA-binding profiles (Wingender et al., 2000). While made accessible on the WWW, it is maintained internally as a relational database management system. First developed as an independent relational database system, we developed PathoDB, a complementary database on pathologically relevant mutations in regulatory regions (promoters, enhancers, etc.) and transcription factor genes. Another genomic characteristic that is important to gene regulation are scaffold / matrix attached regions (S/MARs). Locations, properties and sequences of these regions as well as the proteins interacting with them have been collected in the S/MARt DB (S/MAR transactions database), which like TRANSFAC is a relational database but is provided to the public as flatfiles (Liebich et al., in preparation). These three databases have been integrated in a first step as a SQL database system under MS SQL server.

In its FACTOR table, the TRANSFAC database harbors information about local and global structural features, functional properties including information how the factor activities are regulated, and gene expression patterns. However, the latter information is presented in free text fields only. To provide a more structured representation of this kind of data, we have developed the CYTOMER® database on human and mouse organs/tissues, cell types, physiological systems and their developmental stages (Chen et al., 1999; Wingender et al., 2000). To provide a catalogue of conditional terms, however, we made use of the mechanisms modeled in the TRANSPATH database on signal transduction pathways (Schacherer et al., submitted to BGRS 2000). This database provides information about extracellular triggers of signaling cascades as well as semantic and/or mechanistic views on these pathways, which in most cases extend to complex regulatory intracellular networks. To achieve optimal prerequisites for the automatic construction of pathways, TRANSPATH has been modeled as an object-oriented database. To facilitate smooth integration with the other databases, TRANSPATH has been re-modeled as a relational system. For this purpose, an adapter has been developed that mirrors the OODBMS into a RDBMS, but also allows the reverse process to enable future updating done with the RDBMS without loosing irreversibly the advantages of the OODBMS.

Now, full integration of both TRANSFAC and TRANSPATH allows to model complete signaling pathways starting with extracellular messenger molecules such as hormones, growth factors or others, going through all the steps involving receptors, different kinds of linking proteins, kinases, targeting at defined groups of transcription factors and ending up with sets of proven and/or suggested target genes. Since this system comprises the complete cascades, it enables also to span arbitrary parts of these pathways as "black boxes" whenever our knowledge is incomplete, but the available data are relevant enough to be stored in the database.

Moreover, this comprehensive system allows us to model regulatory intra- as well as intercellular networks. Intracellular networks arise from the numerous cross-talks between distinct signaling pathways, some components of which may operate in a cell-specific manner. Top consider the latter view, the integration with the CYTOMER database provides us with the required framework for modeling. Also, both types of networks appear when genes are regulated in response to certain signal transduction pathways that by themselves encode regulatory components.

The main problem here is the dynamic representation of these molecular data which have actually a static character. Another important aspect of this problem is that such modeling should finally be quantitative. Different models and simulation approaches are developed, but the problem is still far from being solved.

To start our systematic work in this direction, we focus on kinetic modeling signal transduction pathways. We do that by representing the pathways as a complex graph and using the method of kinetic graphs.

Discussion

We have established a comprehensive data resource on different aspects of gene regulation. Starting from a number of conceptionally linked, but physically separate databases, some of them using different database architectures, we integrated them in an overall relational schema. The final data resource enables now to model complete pathways, starting with extracellular messengers (hormones, growth factors, cytokines etc.) and ending up with sets of up- or down-regulated target genes. This will provide an appropriate platform for simulation of these pathways and networks. Moreover, since PathoDB is part of the whole system, the influence of pathological genetic aberrations can be included in these simulations as well.

Acknowledgements

This work has been supported by grants of the German Ministry of Education, Science, Research and Technology (BMBF; 01 KW 9629/7 and 01 KW 9906/1).

References

1. Chen,X., Dress,A., Karas,H., Reuter,I. and Wingender,E. (1999) A database framework for mapping expression patterns. Computer Science and Biology. Proceedings of the German Conference on Bioinformatics (GCB '99). R. Giegerich, R. Hofestädt, T. Lengauer, W. Mewes, D. Schomburg, M. Vingron and E. Wingender (eds.). GBF-Braunschweig and University of Bielefeld, pp. 174-178.
2. Higo,K., Ugawa,Y., Iwamoto,M. and Korenaga,T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database:1999. Nucleic Acids Res., 27, 297-300.
3. Knüppel,R., Dietze,P., Lehnberg,W., Frech,K. and Wingender, E. (1994) TRANSFAC retrieval program: a network model database of eukaryotic transcription regulating sequences and proteins. J. Comput. Biol., 1, 191-198.
4. Kolchanov, N.A., Podkolodnaya, O.A., Ananko, E.A., Ignatieva, E.V., Stepanenko, I.L., Kel-Margoulis, O.V., Kel, A.E., Merkulova, T.I., Goryachkovskaya, T.N., Busygina, T.V., Kolpakov, F.A., Podkolodny, N.L., Naumochkin, A.N., Korostishevskaya, I.M., Romashchenko, A.G. and Overton,G.C. (2000) Transcription Regulatory Regions Database (TRRD): its status in 2000. Nucleic Acids Res., **28**, 298-301.
5. Rombauts,S., Dehais,P., van Montagu,M. and Rouze,P. (1999) PlantCARE, a plant cis-acting regulatory element database. Nucleic Acids Res., 27, 295-296.
6. Volkenstein,M.V., Goldstein,B.N. (1966) A new method for solving the problems of the stationary kinetics of enzymological reaction. Biochim. Biophys. Acta, 115, 471-477.
7. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Pruss,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res., **28**, 316-319.

SOFTWARE AUTOMATED PACKAGE FOR ANALYZING THE DYNAMICS OF CONTROL GENE NETWORKS

Galimzyanov A.V.

Institute of Biology, Ufa Research Center, Russian Academy of Sciences, Ufa, Russia
e-mail: tchuraev@anrb.ru
*Corresponding author

Keywords: control gene networks, computer modeling, object-oriented programming, computer experiments

Resume

Motivation:

Control gene networks (CGN) are structured most intricately and have a variety of possible behavior regimes, leading to considerable problems in studies of their properties on the level of analytical constructions exclusively. For this reason it seems necessary to invoke some additional investigation techniques, particularly computer modeling and computer experiments.

Results:

The present program complex realizes the method of generalized threshold models (GTM) for the analysis of the dynamics of the molecular genetic control systems (MGCS) and is intended to analyze the eukaryotic CGN dynamics with regard to cellular divisions and phases of a cellular cycle.

Availability:

This application software is available on request from the author.

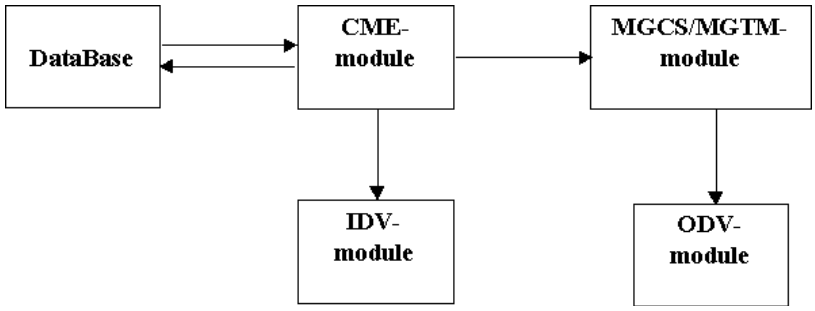


Figure 1. Main components of the program and the diagram of their interactions.

Introduction

The CGN computer modeling expects the solution of two relatively independent problems. On the one hand, one should formulate the description of CGN as a dynamic system made up of many interacting elements, which are different by their nature and occupy different levels of the internal hierarchy, and the interactions themselves (regulatory bonds). On the other hand, this is the program realization of mathematical methods to study the dynamics of CGN functioning on qualitative and quantitative levels. In this work the solution of both formulated problems are presented.

Program structure

The system comprises five principal modules (Fig. 1): CME, DataBase, MGCS/MGTM, IDV and ODV.

1. The CME-module supports the construction and modification environment for CGN models with an arbitrary complexity.
2. The IDV-module provides a visual representation of the characteristics of the CGN model.
3. In the MGCS/MGTM-module the GTM method has been directly realized, and it allows to plot the diagrams of gene activities, kinetic curves for concentrations of gene products (mRNA and proteins), “phase portraits“ in the plane (for two genes), and exact values of respective variables.
4. The ODV-module provides a visual representation of output data in the graphical and tabulated forms.
5. The DataBase-module is meant to store computer experimental data (with the required set of operations: insertion, deletion, extraction and automatic allocation of an element like the CGN model in the program environment).

Methods and algorithms

The program has been developed in terms of object-oriented programming based on peculiar features of the problems on portrait modeling of the behavior of biological systems. The concept of describing an MGCS-type object is based on the gene network mathematical model in the GTM method [1]. With the aim to retain a biological meaning of actual MGCS properties at the level of program realization a direct correlation is

established between the elements and functions of the real MGCS and the components of data structures it is described with. To accomplish this a correlation is made among the objects from different conceptual, logical and functional systems. Thus, MGCS genes are correlated with genetic blocks of its mathematical model; on the basis of the description of the model elements some abstract classes of program realization are elaborated ("MGCS", "Genetic Block", "Genetic Block Control System", "Genetic Block Controlled System", "Regulatory Zone", "Regulatory Interaction"), which classes are a combination of fields, methods and properties reflecting the structure and functional association of the MGCS elements. The arrangement of interaction procedures among the objects with the specific types of abstract classes in question makes it possible to model the behavior of arbitrary MGCS more or less complex (MGCS/MGTM-module). In the DataBase-module the representation is based on tables in the database format "Paradox 7.0" ("MGCS", "Gene", "Site") and their interrelations. The "MGCS" table stores the data of computer experiments on the models; structures of the "Gene" и "Site" tables agree with the fields of the lowest hierarchic levels of the "Regulatory Interaction" and "Genetic Block" classes respectively. In the CEM-module the CGN model is realized through the bidirectional self-connected list, thus making it possible to "collect" its structure in the dynamic manner from an arbitrary number of standard-type elements ("Genetic Block" and "Regulatory Relation").

Implementation and results

With the aid of the developed software an attempt has been made to analyze the dynamics of control subsystem for morphogenesis of Arabidopsis thaliana flower [2], and also the efficiency of the GTM method has been estimated as compared with a specific algorithm [3]. Preliminary data have been got on the dynamics of the system controlling the plane development of the Dr. melanogaster plant. The proposed approach is also good in analyzing the dynamics of other CGN.

References

1. Tchuraev, R. N. (1991). A New Method for the Analysis of the Dynamics of the Molecular Genetic Control Systems. I. Description of the Method of Generalized Threshold Models. J. Theor. Biol., 151, 71-87.
2. Tchuraev, R.N., Galimzyanov, A.V. (2000). Application of the Method of Generalized Threshold Models for the Analysis of the Eucariotical Control Gene Subnetworks. In this Proceeding.
3. Mendoza, L., Alvarrez-Buylla, E.R. (1998). Dynamics of the Genetic Regulatory Network for Arabidopsis thaliana Flower Morphogenesis. J. Theor. Biol., 193, 307-319.

GENE NETWORK ON STORAGE MOBILIZATION IN SEED

Axenovich A.V., *Goryachkovsky T.N., Ananko E.A., Omelyanchuk N.A., Stepanenko I.L.

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

Keywords: gene networks, gene expression regulation, databases

Resume

The process of storage's mobilization in specialized tissues of plant seeds under germination is described in the GeneNet database format in the work presented. GeneNet functioning that controls this stage of ontogenesis in plants is in considerable dependence on environmental impacts and its activation is irreversible. An embryo is provided with germination energy as a result of this GeneNet functioning. Fast development of the process is attained by the presence of regulatory contour with positive feedback: embryo nutrition – synthesis of hibberllines - enzyme activation - embryo nutrition.

Availability:

http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/

Introduction

Higher plants accumulate nutrition storage of proteins, starch, lipids, and phytin in specialized tissues, i.e., endosperm in monocots and cotyledonous in dicots. A germinating embryo makes use of these substances to grow and develop at the initial stages of ontogenesis. More and more publications to highlight molecular mechanisms as the base of germination processes are being issued. TRRD and GeneNet databases [Kolchanov N.A. et al., 2000, Kolpakov F.A. et al., 1998] are rational tools to systematize and analyze the information. The data on regulation of plant germination processes and nutrition of an embryo in development is stored in the GeneNet database and generated by the software program GeneNet Viewer [Kolpakov F.A. and Ananko E.A., 1999] in the form of 3 diagrams: 1) mobilization of hydrocarbons; 2) proteins; 3) fats and phosphates.

Results

A chain of events from seed swelling till the appearance of a seedling or primary roots is adopted to be called germination. Storage of nutrients like proteins, starch, lipids and phytin is accumulated in specialized tissues, i.e., endosperm of monocots and cotyledonous of dicots. Storage of substances occurs in the process of seed maturing in a maternal plant. Prior germination, hydrolytic enzymes (amylases AAMY, AMY32D, AMY3B and proteases EBP1, EBP2) splitting storage substances are inactive and their genes are not expressed (Fig 1).

At the initiation stage, the gene network functioning is inhibited by ABA (Abscisic Acid) phytohormone. Its activation is switched on by an external impact like exposure to water. This impact should be sufficient to cause critical swelling values that vary in different plant species. Penetration of water from outer environment leads to decrease in ABA concentration up to the threshold level. After reaching critical values, return of a seed to a dormancy state is impossible. Thus, by annulling the inhibitory effect of ABA, irreversible process of seed germination is launched. Hydrolytic enzymes having been synthesized earlier are being activated. Penetration of membranes of storage vacuoles changes and hydrolases obtain an access to storage polymers. Cassette activation of genes encoding hydrolytic enzymes (AAMY, AMY32B, EBP, AEP) and that of transcription factors (Gamyb) participating in activation of hydrolytic enzymes occurs. Active enzymes split storage substances providing an embryo's nutrition. Embryonic cells synthesize gibberellins which, in their turn, activate the genes coding the enzymes splitting storages [Skriver K. et al.,1991; Hooley R., 1994]. It leads to increase in concentration of active enzymes and intensifies an embryo's nutrition. The development of seed storage mobilization (SSM) is attained by the basic regulatory contour with the positive feedback: synthesis of gibberellins by growing embryonic cells and GA transport to aleurone and endosperm in monocots or activation of bound inactive GA in cotyledonous of dicots. Besides, the activation of genes encoding hydrolytic enzymes, Ca++ transport, and calmodullin is intensified. Membrane penetration changes and basic hydrolases become stable [Hooley R., 1994]. The access of hydrolase is achieved to storage polymers, i.e., starch, proteins, lipids, and phytin. It contributes to intensification of nutrition in growing embryonic cells and of the gibberellin synthesis by these cells and by scutelum. Gibberellin synthesis activates further mobilization of storages. An excessive concentration of glucose derived from starch hydrolysis inhibits amylase gene (Amy3D) transcription. Thus, the

sugar transport to an embryo is controlled. Negative feedback stabilizes the parameters of the network under sugar over-reproduction.

Water is an environmental impact initiating a gene network functioning. ABA concentration in storing tissues of a seed serves as an inhibitory factor conserving the system in inactive state, whereas gibberellin synthesis by developing embryonic cells amplifies the signal. The gene network is activated by a positive feed back: gibberellin synthesis by an embryo's developing cells leads to activation of genes encoding enzymes and intensifies nutrition of an embryo. Gene network is stabilized by the negative feedback: an excess of glucose inhibits the gene encoding an enzyme necessary for glucose synthesis. Switching off the SSM gene network happens only due to depletion of its components.

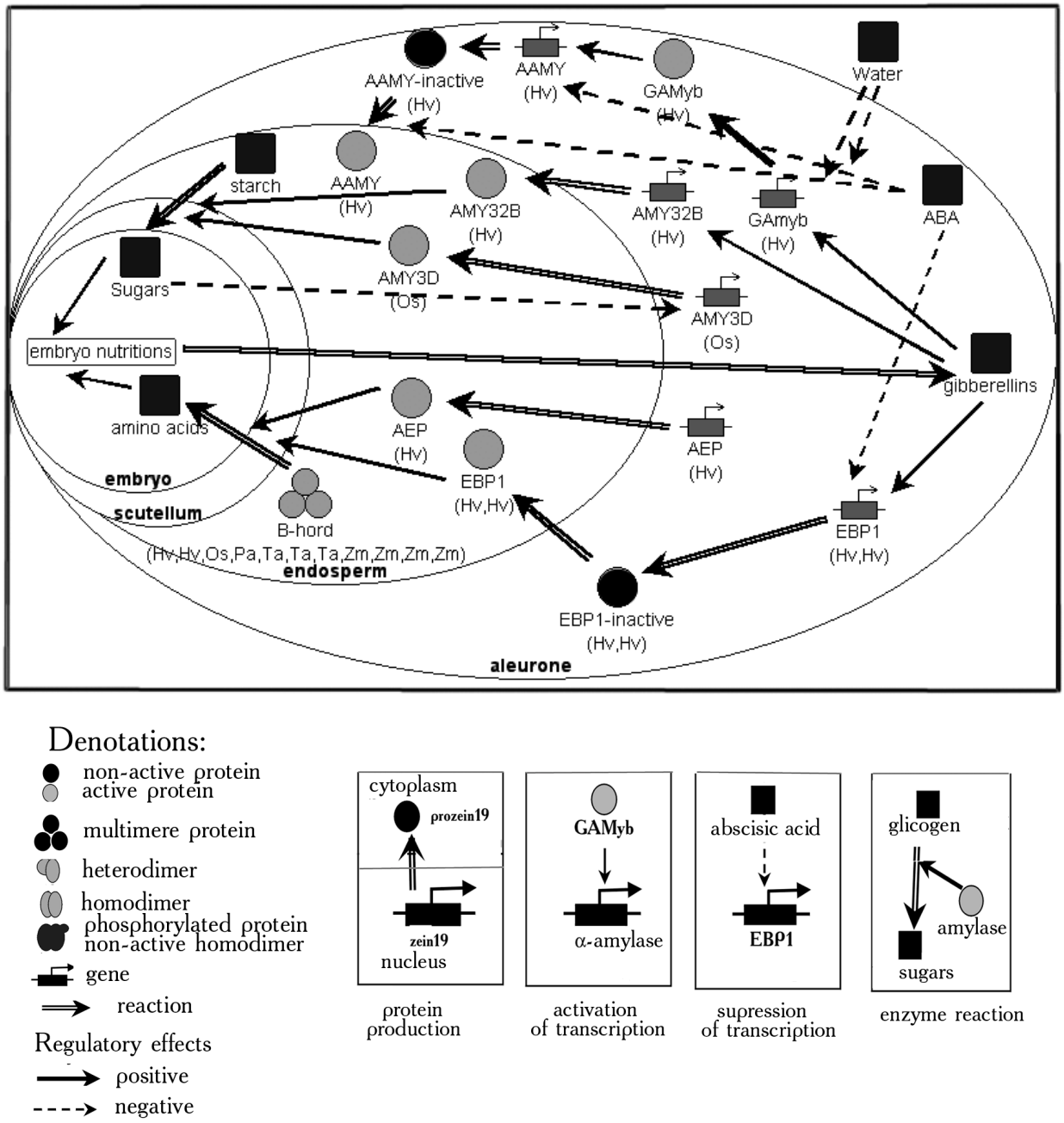


Figure 1. Gene network on seed storage mobilization during germination of grasses. The level of an organism. Rapid and irreversible development of a process is achieved due to regulatory contour with the positive feedback (gibberellin synthesis by the growing plant embryo's cells causes activation of genes encoding enzymes and intensifies embryo' nutrition).

Discussion

The gene networks conception allows considering a separate phenotypic feature, separate process, stage of development, and ontogenesis in a whole as a result of functioning of gene networks, which are organized hierarchically. Every stage of ontogenesis can be figured out as a result of interaction between gene networks. Elementary events occurring at the cellular level play a considerable role. One and the same gene networks function synchronically in all cells of a certain tissue. The presence of regulatory contours supports a definite functional state of a gene network, i.e., timely transport of nutrients into developing embryo) or its transformation into another regime of functioning, which may occur under the action of environmental factors (e.g., change of dominant storage protein under nitrogen deficiency). The action of external signals turns one gene network off and turns on the others. Hormones, Ca++ ions and other low-molecular compounds are molecular carriers of the signals. A gene network with a higher range provides the coordination of this process. A gene network on biosynthesis of storage proteins in plant seeds is under the control of the gene network supporting development of viable seeds. This gene network, in its turn, gets on after a flower's fertilization and is governed by a gene network of a maternal plant having initiated a generative phase of development (Fig 2).

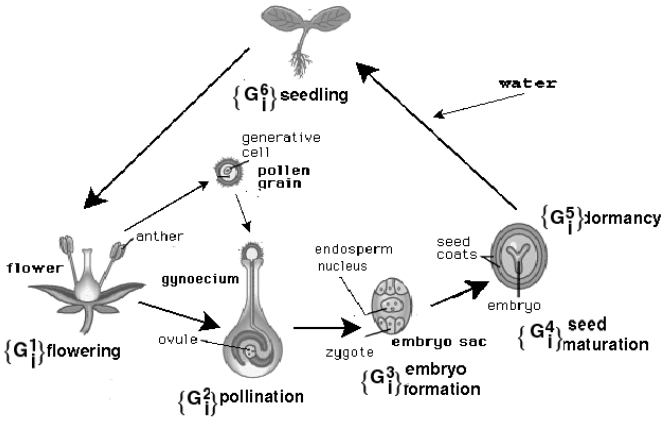


Figure 2. Stage of an ontogenesis of a plant. Each developmental stage corresponds to a set of gene networks {Gik}. The structure of a gene network Gik is formed at the previous stage and is activated by specific factors (fertilization, water).

Higher plants undergo a dormancy stage in their development. At this stage, gene expression is practically absent, gene networks do not function and are repressed. Starting of a novel ontogenetic stage is related to the usage of the storages. Since a plant does not possess by the organs and tissues necessary for photosynthesis at this stage, all the energy needed for germination is taken from the storing tissues. So, the gene network on seed storage mobilization is an example of the highest hierarchical level in the management of ontogenesis. In this gene network, the managing signals are directed not only from the higher levels to the lower ones but vise versa, i.e., gibberellin transport from the tissues of an embryo. The elements of a gene network organizing relationships between the organs and tissues of a plant are located quite distantly: gibberellin synthesis occurs in embryonic tissues, whereas reception of a signal takes place in endosperm.

Acknowledgements

The work is supported by the Russian Foundation for Basic Research (grant No 00-04-49255).

References

1. Hooley R. (1994) Gibberellins: perception, transduction and responses. Plant Mol Biol. Dec;26(5):1529-55.
2. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., Overton G.C. Transcription Regulatory Regions Database (TRRD): its status in 2000 // Nucleic Acids Res. 2000, V. 28, P. 298-301.
3. Kolpakov F.A.,Ananko E.A.,Kolesov G.B.,Kolchanov N.A.(1998). GeneNet: a genenetwork database and its automated visualisation. Bioinformatics, v.14, no.6, P.529-537.
4. Kolpakov F.A. and Ananko E.A. Interactive data input into the GeneNet database (1999) // Bioinformatics Jul-Aug;15(7-8):713-4
5. Skriver K., Olsen F.L., Rogers J.C., Mundy J. Cis-acting DNA elements responsive to gibberellin and its antagonist abscisic acid// Proc.Natl.Acad.Sci.USA. 1991. V.88. P. 7266-7270

SEED MATURATION IN HIGHER PLANTS: GENE NETWORKS ON ONTOGENESIS IN STORAGE TISSUES

Goryachkovsky T.N., Ananko E.A., Kolpakov F.A., Stepanenko I.L.

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

Keywords: gene networks, gene expression regulation, databases

Resume

The work presented is devoted to description in terms of the GeneNet database of a gene networks on ontogenesis in specialized tissues of plant seeds. Among obligatory stages of seed maturation are accumulation and packaging of storages that will be necessary for the germinating embryo. At the initial stage, a considerable increase of a seed in size takes place, mainly due to the storage tissues growth (endosperm in monocots and cotyledonous in dicots). The next stage of seed development is characterized by preparation to dormancy stage. During this period, the so-called LEA (Late Embryogenesis Abundant) genes are being expressed, which preserve the cell components from damaging under dehydration and freezing, together with the genes guarding against premature vivipary of an embryo.

Availability:

Gene networks are described in the format of the GeneNet database [Kolpakov F.A. et al., 1998] and are available via the Internet by the address http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/

Information about each gene, its expression and transcription regulation is stored in the database on regulatory gene regions in eukaryotes TRRD (Transcription Regulatory Region Database) [Kolchanov N.A. et al., 2000], in the section PLANT-TRRD: http://wwwmgs.bionet.nsc.ru/mgs/papers/goryachkovsky/plant-trrd/

Introduction (biology of a process)

The majority of higher plants derive energy for growth and development solely from photosynthesis (autotrophic nutrition). However, each plant undergoes in its ontogenesis a short but important stage of heterotrophic nutrition. The germinating embryo prior appearance of the first photosynthetic structures is fed exclusively by storages accumulated during the seed formation (i.e., proteins, fats, and carbohydrates) [West M.A.L. and Harada J.J. 1993, Shewry P.R., 1995]. As is well known, the main function of a plant seed is to provide a viable offspring. The storage substances play a key role, by providing nutrition of a seedling during the heterotrophic stages of its development. Seed formation passes several overlapping but rather independent stages. Before the beginning of flowering, an embryonic structures are formed out of apical meristem. The switch from the vegetative stage of development to the reproductive one is regulated by different mechanisms. Among these mechanisms are inner biological clock, hormonal background, and environmental conditions such as light day length, temperature fluctuations, humidity, etc.

1. At the first stage, a pollination followed by fertilization occur.
2. The second stage is characterized by establishment of a general composition of the future plant. Embryonic tissues (protoderm, procambium, ground meristem) are being differentiated. The axis of development of an embryo is being formed, with apical meristem of a root from one side and apical meristem of a stem - from the other.
3. At the following stage, storage substances are being intensively worked out, these substances will be necessary at subsequent stages, during seed germination.
4. Seed development is finished by preparation of a seed to the dormancy stage and final maturation at the dormancy stage.
5. Dormancy stage.

Currently, more and more studies are devoted to molecular-genetic events occurring at the cell level and making a foundation for formation of morphological structures. Gene expression regulation is provided within the frames of complex systems, gene networks, which include coordinately functioning genes, the products of these genes expression, and different external signals (hormones, metabolites, etc.).

Gene network on accumulation of storage proteins in endosperm of grasses

Fig. 1 illustrates the gene network controlling biosynthesis of storage proteins during seed maturation. One of the main functions of a seed is to accumulate the nutrition substances (in particular, storage proteins), which will be further utilized by an embryo during germination. The gene network given in Fig. 1. describes the process at the level of an endosperm cell. There are three compartments in endosperm cell: nucleus, cytoplasm, and storage vacuole. Actually, only the genes encoding storage proteins and their activators are expressed in the

cells of storage tissues during this period. Regulation of expression is mainly produced at the level of transcription. Let us consider in more details the elements of the gene network: components and relations between them.

Elements of a gene network:

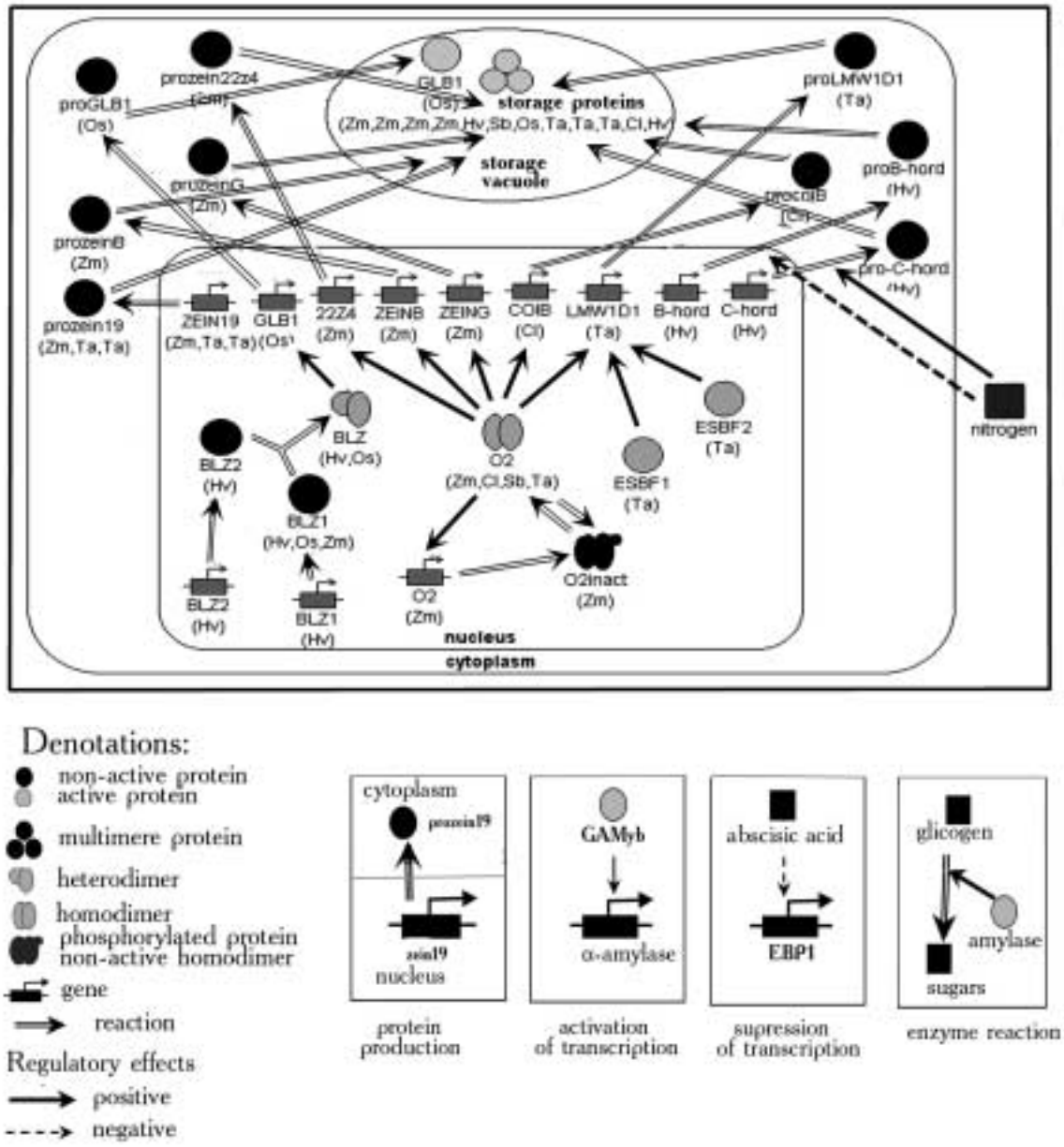


Figure1. A gene network on storage proteins biosynthesis. The level of an endosperm cell in grasses. At this level, only the genes of storage proteins and their activators are expressed in the cells of storing tissues.

1. *Genes encoding storage proteins.* In the gene network, the following genes encoding storage proteins are represented: 22Z4, ZEINB, ZEING , ZEIN19 - zein (Zea mais), b-hord, c-hord - hordein (Hordeum vulgare), Glu-B1 - glutelin (Oriza sativa), PG5a – prolamin, GLB1 - globulin (Oriza sativa), LMW1D1 - glutenin (Triticum aestivum), COIB – coixin (Coix lacryma jobi), KAFA – kafirin (Sorghum bicolor).
2. *Storage proteins.* These proteins are synthesized as monomer subunits that are accumulated in storage vacuoles and being packed into complex subunit structures. Storage proteins subsequently serve as the sources of nitrogen, sulphur, and carbon for the germinating embryo.
3. *Transcription factors.* Transcription factors (O2, BLZ, BLZ1, BLZ2, ESBF1, ESBF2) accumulated at the previous stages or synthesized *de novo* activate the genes encoding storage proteins.
4. *Nitrogen.* Nitrogen is an essential element of storage protein biosynthesis pathway. Its availability in the media influences the gene network functioning.
5. *Relations between the gene network components.* Relations between components of a network are designated in the scheme by arrows. Denotations to the figures are as those given for Fig. 1.

In the GeneNe database, each element of the scheme is supplied by detailed textual comments with references to the TRRD and MEDLINE databases.

Regulatory contours of a gene network:

1. *Positive autoregulation.* Transcription factor Opaque2 (O2) is described in some species of grasses and this factor is a conserved activator of genes involved in a gene network considered. Activation of O2 occurs as a result of dephosphorylation. Enhancement of transcription rate is produced by positive auto-regulation (Fig.1). In promoter region of O2 gene, there is a site recognized by O2. In such a way, self-activation of this gene takes place.
2. *Action of an environment.* The content of the storage protein, to some extent, depends upon availability of initial elements in environment. For example, under conditions of nitrogen excess in an environment, the hordein C is primarily expressed. In the norm, the dominating fraction is hordein B.

Regulatory mechanisms

Functional analysis of promoter regions of embryo-specific genes has revealed the conserved transcription factor binding sites (Table 1).

Table 1. Regulatory elements of embryo-specific genes.

Regulatory element	Sequence	Binding protein	Number in the TRRD database
Endosperm motif Em	TGTAAAAGT	ESBF	12
GCN4-like motif	ATGA(C/G)TCAT	SPA	10
CAAT box	CAAT	unknown	5
O2 binding site	CACGTC	Opaque 2	3
G box	ACGT	GBFs	2
AACA motif	CAACAAA	unknown	6
Legumin box	CATGCAT	unknown	3

Within the region in-between -300 bp from transcription start site, an endosperm-box is localized within many genes encoding storage proteins. This element is necessary for producing maximal level of storage protein gene expression in endosperm of grasses. Endosperm-box is represented by a regulatory element compiled out of two synergetic motifs: Em and GCN4-like motif. The GCN4-like motif in combination with the neighboring endosperm-motif participate in regulation of expression of the C-hordein gene, encoding dominating storage protein in barley, in response to alterations of nitrogen concentration in environment (Fig. 2). The low nitrogen concentration in the media may invert the functions of the GCN4 element: it becomes negative, then D-hordein dominates in barley endosperm.

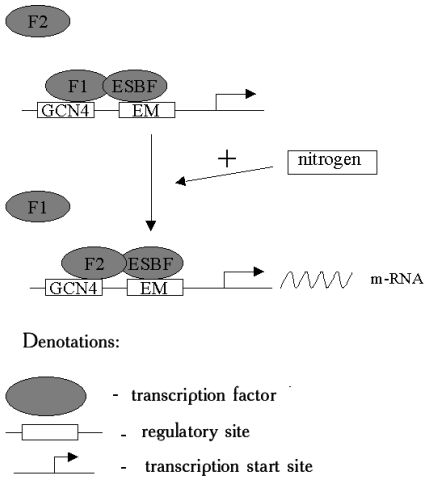


Figure 2. C hordein gene expression. Endosperm box binding factor (ESBF) binds to the endosperm-motif Em of the C hordein gene and interacts with hypothetical factors F1 and F2. Under conditions of nitrogen deficient, GCN4-motif interacts with the F1 factor and the gene transcription is suppressed. When nitrogen incomes from the environment, the GCN4-motif interacts with the F2 transcription factor, thus, causing transcription activation.

The gene network on storage proteins biosynthesis provides a transition of a seed from the stage of main embryonic structures formation to almost mature state. After accumulating of a sufficient amount of resources, the gene network considered stops its functioning and the genes of storage proteins biosynthesis are repressed. The mechanism terminating an accumulation of storage substances is still unknown.

Gene network on preparation of a seed to a dormancy period

In Fig. 3, the gene network is demonstrated, which gives a description of the subsequent stage of a seed maturation, also at the level of an endosperm cell. Switching on of a gene network causes termination of tissue growth, seed dehydration and transition to the dormancy stage. Transition to this stage is related to considerable accumulation of abscisic acid (ABA) in seed tissues. An endosperm cell in this case contains four compartments: nucleus, cytoplasm, storage vacuole, and oleosome. The gene network includes the following components and regulatory contours:

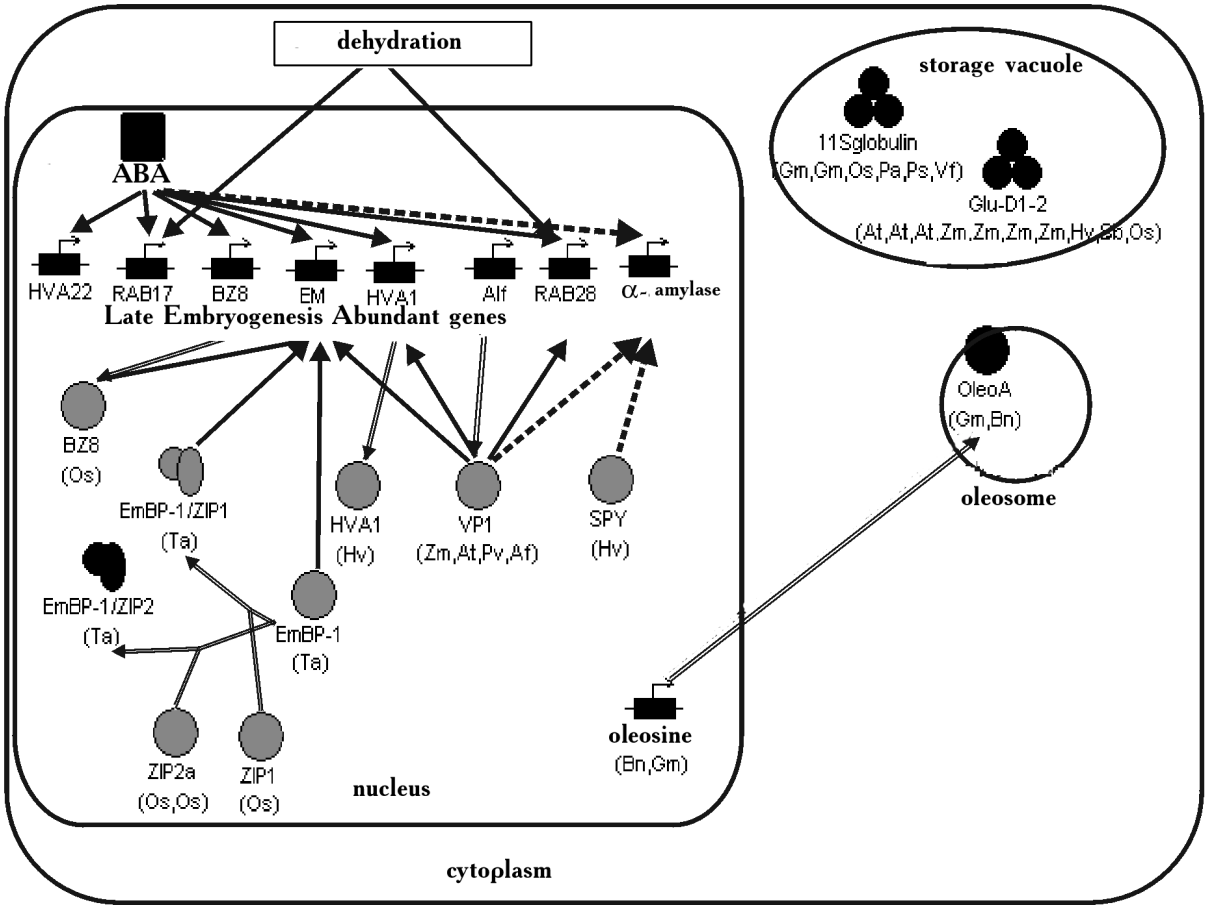


Figure 3. A gene network on a seed preparation for dormation period. Denotations are as in Fig. 1. Abscisic acid activates the genes at the late stages of embryogenesis. As a result, an embryo becomes stabilized by dehydration and it is able to be in a dormancy state for a durable time.

Components of a gene network:

1. *Late embryogenesis abundant genes (LEA).* An important element of ontogenesis program realization that follows accumulation of storage substances is the Em gene. Its transcription is induced by ABA and is enhanced by EmBP-1 and Opaque 2 transcription factors. The HVA22, HVA1 (Hordeum vulgare) gene products protect the seed tissues from damaging under dehydration in the process of preparation to dormancy. Responsive to ABA genes (RAB17 and RAB28) participate in the signal transduction from ABA. Oleosines perform a structural function in oleosome's membranes.
2. *Genes encoding hydrolytic enzymes* (α-amilase) are repressed at this stage.
3. *Hormones.* The phytohormone ABA is synthesized in plants in response to various stress factors action. In the gene network considered, the ABA inhibits premature vivipary of seeds and induces a considerable bulk of the known LEA genes.
4. *Transcription factors* (BLZ8, EmBP-1, ZIP1, ZIP2, ZIP2a, VP1, SPY) govern the coordinated functioning of a gene network. Vp1 (Zea mays) is a conserved element of the ABA signal transduction pathway. Its analogs are known in *Arabidopsis* (Abi-3), in beans (PvAlf), and in oats (AfVP1).
5. *Relations between components of a gene network* are denoted analogously to those represented in Fig. 1.

Regulatory contours:

1. *Positive feedback* as in the previous case stimulates the process. As the result of the gene net functioning, a dehydration of seeds takes place, which in its turn induces the genes of late embryogenesis and the products of these genes provide further dehydration.
2. *Processes*. As the result of a gene network functioning, the following processes occur: dehydration of seeds; accumulation of fat storages; strengthening of cell walls; inhibition of activity of almost all genes.

The key regulator of the second gene network, which inhibits premature vivipary of seeds and induces the most part of the known LEA genes, is a phytohormone ABA, abscisic acid. The product of the Vp1 gene mediates the signal transduction. The exact mechanism of seed dehydration is unknown to date.

At this stage of maturation, the genes encoding oleosines, which are the structure proteins of an oleosome membrane, are being expressed. Oleosomes accumulate storages of fats. The surface of these organelles is represented by a phospholipid monolayer with integrated molecules of oleosine. 5'-regulatory region of the soybean oleosine contains similar regulatory elements to those of soybean storage proteins and LEA genes.

Thus, development of a plant suspends, an embryo is stabilized by dehydration and may stay in the dormancy stage for a long period of time.

Acknowledgements

The work is supported by the Russian Foundation for Basic Research (grant No 00-04-49255). The authors are grateful to G. Orlova for the help in translation of the manuscript.

References

1. Kolchanov N.A., Podkolodnaya O.A., Ananko E.A., Ignatieva E.V., Stepanenko I.L., Kel-Margoulis O.V., Kel A.E., Merkulova T.I., Goryachkovskaya T.N., Busygina T.V., Kolpakov F.A., Podkolodny N.L., Naumochkin A.N., Korostishevskaya I.M., Romashchenko A.G., Overton G.C. Transcription Regulatory Regions Database (TRRD): its status in 2000 // Nucleic Acids Res. 2000, V. 28, P. 298-301.
2. Kolpakov F.A., Ananko E.A. Kolesov G.B. Kolchanov N.A. GeneNet: a gene network database and its automated visualization, Bioinformatics, 1998, 14, 529 – 537.
3. Shewry P.R. Plant storage proteins. Biol Rev Camb Philos Soc. 1995 70(3):375-426.
4. West M.A.L., Harada J.J., 1993, Embryogenesis in higher plants. An overview. Plant Cell, 5, 1361-1369

NUMERICAL STUDY OF MATHEMATICAL MODELS DESCRIBED DYNAMICS OF GENE NETS FUNCTIONING: SOFTWARE PACKAGE STEP

[1]Berezin A.Yu., [1]Gainoval.A., [2]MatushkinYu.G.,[2]Likhoshvai V.A., [1]Fadeev S.I.

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

Keywords: mathematical model, numerical analysis, gene networks

Resume

Motivation:

The models based on the concept about the chemical and kinetical nature of the principal molecular processes in the cell are an important class of mathematical models describing dynamics of gene network functioning. As a rule, such models belong to the class of autonomous equation systems. An important stage of investigation of such models is determination of key characteristics of their behavior for appropriate model parameters. It is necessary to apply the specific methods and techniques.

Results:

In this paper, we describe the software package STEP designed for the complex study of nonlinear equations systems and autonomous systems of general type involving the "standard" set of computational problems like the stability of stationary solutions, determination of oscillations, investigation of stationary solutions (i.e., solutions to nonlinear equations systems) by a parameter, multiplicity of solutions.

Problem of numerical study of autonomous systems is used in many applications such as the mathematical modeling of catalytic reactions, biological processes, physical problems and mechanics. The basic purpose therewith is to detect the nonlinear phenomena described by the mathematical model, namely, detection of the model parameters, characterized by the oscillations, multiplicity of solutions and high parametric sensitivity in their neighborhood, etc.

Thorough parametric analysis of autonomous system behavior including the determination of the stationary solutions is possible only for some models (for example, "brusselator" model). In general, investigation of autonomous systems for N ordinary differential equations

$$dy/dt=f(y,\alpha), \qquad (1)$$

where α is a model parameter, bear a numerical experiment character. The methods not accounting for the specific nature of right-hands of system, allow:

A) to construct numerically the dependence $y(\alpha)$ of stationary solutions to the system

$$f(y,\alpha)=0, \qquad (2)$$

by the method of solution continuation by parameter and to find simultaneously the α-regions of multiple solutions;

B) to determine their stability and locate the points on the stationary solutions diagram where unstable stationary solution passes into a stable limit cycle (Hoph bifurcation);

C) to obtain the oscillations by integrating the autonomous system or to seek of the start stationary solution by a continuation method by parameter α;

D) to detect α-regions , in which all stationary solutions are unstable. For autonomous system, it means the self-excitation of oscillations for arbitrary initial data.

The software package STEP represents a powerful tool for realization of above-mentioned points. This package was developed by research group headed by professor Fadeev. The package STEP is based on the original algorithms suggested in the Sobolev Institute of Mathematics (Novosibirsk), including the method of continuation of stationary solutions on parameter, technique of determination of stationary solution stability and integration of stiff autonomous system.

Brief description of the algorithms

Nonlinear system (2) is supposed to be presented by the smooth spatial curve S and, therefore, only singular points of "turn" type take place there in numerical constructing S. Here the Jacobi matrix $f_y(y,\alpha)$ is degenerated. Since under the assumption that the rank $(N \times (N+1))$-matrix $[f_y(y,\alpha), f_\alpha(y,\alpha)]$ would be always equal to N in the neighborhood of S, then by the implicit function theorem such component μ of combined vector (y,α) could be found as $f_x(x,\mu) \neq 0$, where x is a vector (y,α) without the component μ. It means that the solution to system (2) can be continued to one step by the parameter μ, which will be called a current parameter. If the solution is found for the current parameter λ, then the choice of novel current parameter μ is stipulated by the normalization of the solution derivatives with respect to λ. Let

$| dx_k/d\lambda | = \max (|dy_1/d\lambda|,|dy_2/d\lambda|,...,|dy_N/d\lambda|,|d\alpha/d\lambda|)$.

Then $\mu=x_k$. Herewith, the quantities

$| dy_i/d\mu | = |dy_i/d\lambda| / |dx_k/d\lambda|$, i=1,2,...,N, $|d\alpha/d\mu| = |d\alpha/d\lambda| / |dx_k/d\lambda|$,

are less or equal to 1. Further, the derivatives found with respect to current parameter μ are used for prediction of the initial approximation in the neighborhood of μ by the Newton's method, etc.

We have used a numerical κ-criterion by Godunov-Bulgakov to determine the stability of stationary solutions in the package. It is based on the effective technique of determining the matrix norm H of the solution to Liapunov matrix equation

$HA + A^*H=I$,

where $A = f_y(y,\alpha)$. The STEP package contains a method well suited for the stiff initial value problems as well, namely the semi-implicit Rosenbrock method of the 2 order. For details, see [Fadeev et al, 1998]. List of references points to comprehensive practice of using the STEP package in applications.

Package STEP runs on IBM-compatible personal computers. In order to construct the mathematical model, it is required to input the expressions for right-hands of a system, i.e., the elements of the vector function $f(y,\alpha)$. The elements of matrix $[f_y, f_\alpha]$ are evaluated numerically on the basis of the Richardson approximation. Due to this fact, the volume of input information is considerably decreased, and the model correction related to numerical investigation of system (2) depending on a parameter proceeds with a minimum efforts.

In essence, it is only necessary to find out a model parameter that would play a role of a parameter α. Then it is required to input the elements of the matrix $f_y(y,\alpha)$ for the studying of stability of autonomous system stationary solutions.

Examples of the software package STEP applications

The modified Prigojin's model [Prigojin & Stengers, 1986] is as follows:

$$
\begin{aligned}
dy_1/dt &= k_{13} - k_{11}y_1 + (-k_3y_1y_5+k_4y_3)\\
dy_2/dt &= k_{14} - k_{12}y_2 + (-k_{10}y_2+k_9y_4)\\
dy_3/dt &= -(-k_3y_1y_5+k_4y_3) - (k_5y_3-k_6y_6) \qquad (3)\\
dy_4/dt &= 2(-k_1y_4^2+k_2y_5) + (-k_8y_4y_5+k_7y_6) - (-k_{10}y_2+k_9y_4)\\
dy_5/dt &= -(-k_1y_4^2+k_2y_5) + (-k_8y_4y_5+k_7y_6) + (-k_3y_1y_5+k_4y_3)\\
dy_6/dt &= -(-k_8y_4y_5+k_7y_6) + (k_5y_3-k_6y_6),
\end{aligned}
$$

where k_i, i=1,...,14, are the model parameters. As the parameter α serves k_1.

The stationary solution diagram for the following parameters: $0 < \alpha < 10^4$, $k_2=k_4=k_5=k_6=k_7=k_9=k_{10}=k_{11}=1$, $k_3=0.0001$, $k_8=0.01$, $k_{12}=10$, $k_{13}=100$, $k_{14}=0.1$, is shown in Figure 1. As can be seen from the diagram, there exist α-regions limited by the "turn" points α=1078 and α=6959, such that there exist 1, 3 or 1 stationary solutions, respectively. The change of stability character takes place in the "turn" points: stability $\rightarrow$ instability $\rightarrow$ stability. It gives the evidence about the absence of oscillations as well.

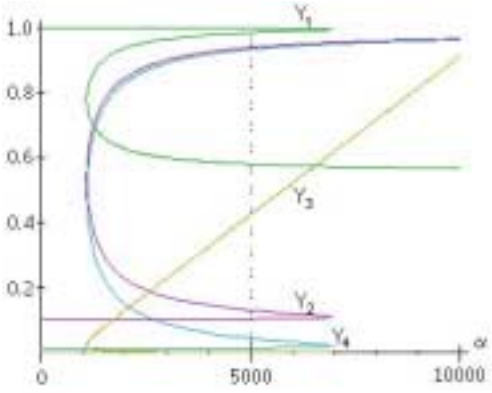


Figure 1. Stationary solutions diagram for Prigogin's model.

Brusselator model [Volkenshtein, 1988]:

$$dx/dt=k_1A+k_2x^2y-k_3Bx-k_4x, \qquad dy/dt=-k_2x^2y+k_3Bx, \qquad (4)$$

where k_i, i=1,...,4, A, B are the parameters. A parameter plays the role of the parameter α. Investigation of stability for k_i=1, i=1,...,4, B=2 shows that if $\alpha < 1$, then the stationary solutions are unstable, and if $\alpha > 1$, then

stationary solutions are stable. Therefore, for arbitrary nonnegative initial conditions, the solution to system (4) proceeds to the oscillatory regime, if $\alpha < 1$ (Fig. 2), or to the stationary ones, if $\alpha > 1$ (Fig. 3).

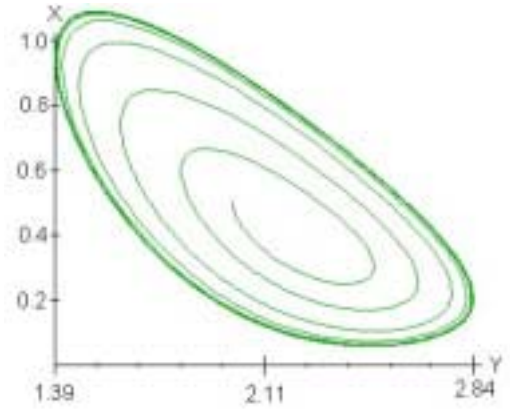


Figure 2. Oscillations for brusselator model

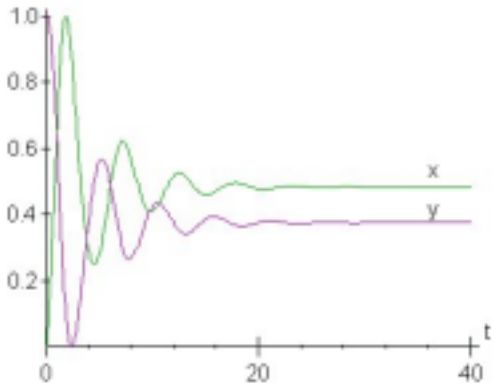


Figure 3. Proceeding to the stationary regime in brusselator model

Acknowledgements

The work was supported by the grant No. 106 of the State R&D Program *Human Genome* and Integrative Project of SB RAS No 66 *Simulation of basic genetic processes and systems*. The authors are grateful to N.A. Kolchanov for helpful discussions.

References

1. S.I.Fadeev, S.A.Pokrovskaja, A.Yu.Berezin, I.A.Gainova, "Software package STEP for numerical investigation of systems of nonlinear equations and autonomous systems of general type. Description of operation of the package STEP on examples of problems from the course "Engineering chemistry of catalytic processes" Novosibirsk State University (1998).
2. I. Prigojin & I. Stengers, "Order from Chaos" M., Progress (1986). (in Rus)
3. M.Volkenshtein, "Biophysics" M., Nauka (1988).

THE RECONSTRUCTION OF THE DROSOPHILA SEGMENTATION MECHANISMS FROM EXPERIMENTAL DATA: PROCESSING AND ANALYSIS OF CONFOCAL IMAGES OF EXPRESSION PATTERNS

[1]Spirov A.V., [2]Timakin D.L., [3]Reinitz J., [3]Kosman D., [1]Spirova O.A.

[1]The Sechenov Institute of Evolutionary Physiology and Biochemistry, 44 Thorez Ave., St. Petersburg, Russia; spirov@iephb.nw.ru;
[2]St.Petersburg State Technical University, Computer Science and Engineering Faculty, System e-mail: dtimakin@mail.ru, dave@eve.molbio.mssm.edu
[3]Dept. of Biochemistry and Molecular Biology, Box 1020 Mt. Sinai Medical School, One Gustave L., USA; reinitz@kruppel.molbio.mssm.edu, spirov@iephb.nw.ru,
*Corresponding author

Keywords: genetic networks, genes of segmentation, pattern of expression, confocal microscopy, image analysis, image processing, computer analysis, reconstruction of gene regulation, atlas of expression, Drosophila melanogaster

Resume

Motivation:

Modern large-scale "functional genomics" projects are inconceivable without the automated processing and computer-aided analysis of images. The project we are engaged in is aimed at the automated transformation of gene expression data in confocally scanned images of the fruit fly (*Drosophila melanogaster*) embryos into an electronic database of expression. For the detailed reconstruction of the fly segmentation mechanisms we need to receive statistically authentic summary picture of detailed pattern dynamics proceeding from a large number of treated embryos.

Results:

Our investigations have leaded us to conclusion about necessity of minimum three procedures of processing of images for compiling of final data set. It is *elastic deformation*, *registration* and *interpolation* procedures. We have developed and tested family of programs for the three image-processing steps. Biologically significant results obtained by these procedures will be discussed.

Availability:

The codes of the programs in C++ and documentation are available via Internet at http://www.mssm.edu/molbio/hoxpro/atlas/atlas.html

Introduction

Computer-Aided Analysis of Biological Images.

The ongoing revolution in molecular genetics has progressed from the large-scale automated characterization of genomic sequence to the characterization of the biological function of the genome. These investigations mark the beginning of the era of "functional genomics" (Lander, 1996). A key feature of genomic scale approaches is the automated treatment of large amounts of data. Both current and future work in the field is impossible without the automated processing and computer-aided analysis of images in connection with updating interactive electronic image databases.

A key aspect of such processing involves the *segmentation* of individual images, the *registration* of serial images, and the *interpolation* of 2D fields of concentrations of segmentation factors. Many problems involving the recognition, classification, segmentation, registration, and interpolation of images can be formulated as optimization problems. Contemporary approaches based on *evolutionary computations* are a promising avenue for the solution of such problems.

The work reported here is part of a large scale project to construct a model of segment determination in the fruit fly *Drosophila melanogaster* based on coarse-grained chemical kinetic equations (Reinitz et al., 1998). The acquisition and mapping of gene expression data at a heretofore unprecedented level of precision is an integral part of this project. The current emphasis in our work is on the automated transformation of gene expression data in confocally scanned images into an electronic database of expression. Here we describe our evolutionary computations-based approach to image processing for quantitative atlas of *Drosophila* genes expression. Biologically significant results obtained by these procedures will be discussed.

Methods and algorithms

Images of Drosophila Genes Expression

Processing of embryonic images begin with data expressed in terms of the average fluorescence level (proportional to gene expression level) of each nucleus, where segmentation proteins exert their biological function. This data was obtained as follows.

Antibodies for 15 protein products of segmentation genes were raised and over 1000 images were prepared and scanned (Kosman and Reinitz, 1998). These images were computationally treated by means of the *Khoros* package. Embryos were rotated and cropped automatically. Next, the images were *segmented* (Kosman et al. in preparation). About 2000-2500 segmented and identified nuclei are obtained from each image. Each nucleus is labeled numerically, and the *x* and *y* coordinates of its centroid are found, together with the average fluorescence level over that nucleus. The segmented data takes the form of tables in ASCII text format. The result is the conversion of an image to a set of numerical data which is then suitable for further processing.

Elastic Deformations: "Stripe Straightening" Procedure

Early in the development the fruit fly embryo is shaped roughly like a hollow prolate ellipsoid, composed of a shell of nuclei which are not separated by cell membranes. Deviations from the ellipsoidal shape reveal the future polarity of the animal's body: The more pointed end on the long axis makes anterior (head) structures, and the rounder end posterior (tail) structures. From a lateral (side) perspective, one long edge of the embryo is flat and will make dorsal ("back"') structures, while the other long edge is rounded and makes ventral ("underside"') structures. But what is more, so called *pair-rule stripes* (early markers of the future segmental pattern (Akam, 1987) are not parallel and straight, but have a crescent-like form. The curvature of the stripes is highest at the termini, and minimal at the central part. Each stripe specifies an anterior-posterior location, and these stripes can be regarded as contours in an *intrinsic coordinate system* (Spirov et al., 2000) that is being created by the embryo itself.

Our data processing begins with a smooth transformation of spatial coordinates. If the image is smoothly transformed such that the curvilinear coordinates are plotted orthogonally, the stripes appear straight, so the determination of these coordinates can be viewed as a "stripe straightening"' procedure (Spirov et al., 2000).

Registration of Serial Images

Our next procedure is registration of serial images. We need this treatment for full-scale quantitative comparisons and analysis of pattern dynamics at a single nucleus resolution. Registration cannot be performed on a direct nucleus by nucleus basis because of individual differences among embryos. Moreover, close inspection of the edge of a well-demarcated expression domain shows irregularity due to the arrangement of nuclei, which do not lie on a rectangular or hexagonal grid. What is more, any two embryos of the same age can differ in size and form.

Our preliminary at hand computations demonstrated that registration of *Drosophila* early blastoderm images takes elastic deformations. So we can use practically the same approach, as is the case of the stripe straightening problem.

Interpolation

Because blastoderm nuclei don't form either regular square or regular hexagonal mesh one-dimensional and two-dimensional interpolation of expression patterns are non-trivial computational problems. However resolution of this problem is essential requirement for the dynamical modeling and statistical analysis of the expression data. Hence all of the following calculations drastically depend on the correct identification of interpolation function for this irregular mesh.

We used and compared several standard approaches for 1D and 2D interpolation. Particularly it was 1D and 2D spline (cubic spline) and 1D and 2D Fourier interpolations. However all these procedures require either regular mesh or are based on transition to a regular mesh. And it has appeared completely unacceptable for the level of precision, which is pursued in our project.

All this has motivated us to take advantage of an interpolation by truncated two-dimensional Fourier polynomials. The power of series was chosen empirically. The Fourier coefficients were found by optimization techniques, while comparison of interpolation result was performed on given irregular mesh of each image under treatment.

Availability

The codes in C++ for the programs for elastic deformations, registration and interpolation, as well appropriate documentation, are available via Internet at http://www.mssm.edu/molbio/hoxpro/atlas/atlas.html.

Results and Discussion.

Experimental Determination of Drosophila Embryonic Coordinates

This elastic deformation procedure called stripe straightening gives possibility to found following:
1. The intrinsic coordinates found for primary pair-rule gene *even-skipped* fits for other gap and pair-rule genes;
2. Curvilinearity of coordinates in anterior and posterior parts of an embryo is controlled by autonomous mechanisms;
3. Curvilinearity of coordinates in anterior part probably is under control of the morphogen *bicoid* gradient.

Statistical Features of Expression of the Segmentation Genes at Single-Nucleus Resolution

It is shown, that
1. The space distribution and the level of a fractional error of fluorescence of the segmentation factors correlates with a position of the factor in the cascade of the genes-controllers of segmentation;
2. In limits of an investigated time interval (from early to late cleavage cycle 14) the irregularity of the boundaries of domains of the genes expression increases;
3. For genes of pair-rule group at least up to late 14 division stage, the evident increase of a proportion of the nuclei is observed, the level of an expression of these genes in which is essentially (~ 5-20 %) differs from average for the nearest neighbors.

Temporal Ranking, Registration and Classification of the Time-Series of Expression Patterns

The developed methods of serial registration of images make possible to realize the following procedures:
1. Automatic ranking of images according to "maturity" of pattern expression resulting in temporally ordered series beginning from earliest up to mature patterns;
2. Automatic subdivision of obtained time series according to known temporal stages (in accordance with accepted classification);
3. Registration of images on the basis of obtained time-series;
4. Visual methods of comparison of pattern dynamics of different genes with high precision;
5. Comparison of stages and rate of the expression patterns maturation for any pairs of studied genes.

Acknowledgements

This work is supported by RFBR, grant No 00-04-48515; INTAS, grant No 97-30950; USA National Institutes of Health, grant RO1-RR07801; and GAP awards RBO-685 and RBO-895.

References

4. Akam,M. (1987) The molecular basis for metameric pattern in the Drosophila embryo. **Development** 101:1-22.
5. Kosman,D. and Reinitz,J. (1998) Rapid preparation of a panel of polyclonal antibodies to Drosophila segmentation proteins**. Development, Genes, and Evolution** 208:290-294.
6. Lander E.S., (1996) The new genomics: Global view of biology, **Science**, 274:536.
7. Reinitz,J., Kosman,D., Vanario-Alonso,C.E. Sharp,D. (1998) Stripe forming architecture of the gap gene system. **Developmental Genetics** 23:11-27.
8. Spirov,A.V., Timakin,D.L., Reinitz,J. and Kosman,D. (2000) Experimental Determination Of Drosophila Embryonic Coordinates By Genetic Algorithms, the Simplex Method, And Their Hybrid. In Proceedings of Second European Workshop On Evolutionary Computation In Image Analysis And Signal Processing, Edinburgh, April 17, 2000.

AUTHOR INDEX

N

O

P

R

S

T

V

W

Y

Z

KEYWORDS INDEX

K

L

M

N

O

P

R

S

T

U

V

W

Y

Z

T