RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

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

Volume 2

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

IC&G, Novosibirsk, 2002

International Program Committee

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

Local Organizing Committee

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

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

Our sponsors

Organizers

Institute of Cytology and Genetics, SB RAS

Siberian Branch of the Russian Academy of Sciences

Grants

INTAS Conference Grant

Glaxo Wellcome Inc.

Russian Foundation for Basic Research

Ministry of Industry, Science and Technologies of the Russian Federation

Information sponsors

San Diego Supercomputer Center, United States

RIKEN Tsukuba Institute

Bielefeld University, Faculty of Technology

http://www.karger.com/

In Silico Biology

Others

KWESTA-group: computers, computer accessories, service

CONTENTS

COMPARATIVE AND EVOLUTIONARY GENOMICS

SYSTEM COMPUTATIONAL BIOLOGY: ANALYSIS AND MODELLING OF GENE NETWORKS AND METABOLIC PATHWAYS

INTRODUCTION

Four volumes of Proceedings of the Third International Conference on Bioinformatics of Genome Regulation and Structure – BGRS' 2002 (Akademgorodok, Novosibirsk, Russia, July 14-20, 2002) incorporate about 180 annotated extended abstracts (short papers) devoted to the actual problems in bioinformatics of genome regulation and structure.

The Conference BGRS' 2002 is organized by the Laboratory of Theoretical Genetics of the Institute of Cytology and Genetics of Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia. BGRS' 2002 is the third in the series. It will continue the traditions of the previous conferences, BGRS' 98 and BGRS' 2000, which were held in Novosibirsk in August 1998 and 2000, respectively.

As the greatest scientific event within the period between the Conferences BGRS' 2000 and BGRS' 2002, could be undoubtedly viewed the completion of human genome draft sequencing. This event has initiated the beginning of the post-genome era in biology. This era is characterized by sharp increase in research scale in the fields of transcriptomics, proteomics, and systemic biology (gene interaction, gene network functioning, signal transduction pathways), without loosing the fundamental interest to studying structural genome organization.

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

The huge volume of experimental data that has been acquired on genome structure, functioning and gene expression regulation demonstrate the blistering growth. 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 four-volume issues of the BGRS' 2002, the Program Committee came to a conclusion that participants of the Conference have concentrated their attention at consideration of the hottest items in bioinformatics listed below: (i) regulatory genomic sequences: databases, knowledge bases, computer analysis, modelling and recognition; (ii) large-scale genome analysis and functional annotation; (iii) gene structure finding and prediction; (iv) comparative and evolutionary genomics; (v) computer analysis of genome polymorphism and evolution; computer analysis and modelling of transcription, splicing and translation; structural computational biology - genomic DNA, RNA and protein structural and functional organization; (vi) gene networks, signal transduction pathways and genetically controlled metabolic pathways: databases, knowledge bases, computer analysis, and modelling; principles of organization, functioning, and evolution (vii) data warehousing, Knowledge Discovery and Data Mining; (viii) analysis of fundamental regularities in genome functioning, organization, and evolution.

The researchers working in the fields of experimental biology are also invited to participate in the work of BGRS' 2002 in order to develop a sort of interface between experimental and computer-assisted researches in the fields of genomics, transcriptomics, proteomics, structural and systemic biology, as well as for contributing to promotion of computational biology to experimental research. These results are highlighted in the fourth volume of BGRS' 2002 Proceedings.

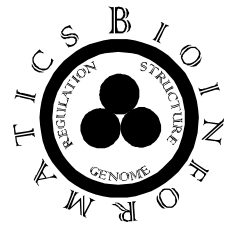
All the questions listed above will be suggested to consideration of participants of BGRS' 2002 at plenary lectures, oral communications, poster sessions, Internet computer demonstrations, and round-table discussions.

The Conference is sponsored by Siberian Branch of the Russian Academy of Sciences, by the Institute of Cytology and Genetics SB RAS, by Russian Foundation for Basic Research, by Russian Ministry of Industry, Science and Technologies, by the Company Glaxo Research and Development Limited, by independent International Association formed by the European Community INTAS. The Organizing Committee of the Conference tender thanks to all the sponsors for financial support.

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

Professor Phil Bourne
SDSC, San-Diego, USA
Co-Chairman of the Conference

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



COMPARATIVE AND EVOLUTIONARY GENOMICS

COMPARATIVE STUDY OF THE ORTHOPOXVIRUS GENES A27L, A56R, B8R, AND C11R

**** Babkin I.V., Mikheev M.V., Shchelkunov S.N.***

State Research Center of Virology and Biotechnology Vector, Koltsovo, Novosibirsk region, 633159, Russia,
e-mail: babkina@vector.nsc.ru
*To whom correspondence should be addressed

***Key words**: orthopoxviruses, genome, evolution, computer analysis*

Abstract

Motivation: Orthopoxviruses are a large group of closely related viruses displaying a high homology of their genomic nucleotide sequences. These viruses differ also in the severities of the diseases they cause and have different host ranges. Evolutionary relationships of various orthopoxvirus species are yet vague (Marennikova, Shchelkunov, 1998).
Results: Analysis of the structures of important virulent factors, such as virus growth factor, gamma-interferon receptor, hemagglutinin, and fusion protein, demonstrated that the corresponding variola virus genes display most numerous distinctions from the other orthopoxviruses, including the functionally important gene regions. In phylogenetic relationships, the genes in question fall into natural groups according to their species attribution. Phylogenetic relationships between the studied orthopoxviruses found while analyzing different genes display certain distinctions. Analysis of the EGF-like region of growth factors demonstrated that orthopoxvirus growth factors clustered with epiregulin, whereas leporipoxvirus growth factors, with transforming growth factors and those of orf viruses, with epidermal growth factors. Presumably, this suggests that different cellular genes from the family of EGF-like growth factors were donated to different poxviruses during their evolutionary development.

Introduction

Poxviruses are the largest, intricately organized animal viruses. The viruses from the genus Orthopoxvirus are most well studied. In particular, this may be explained by the fact that such viruses as variola and vaccinia belong to orthopoxviruses. However, evolutionary relationships of orthopoxviruses are yet unclear. The viruses closely related to variola virus (VAR), such as monkeypox (MPV) and cowpox (CPV) viruses, persist in nature. MPV causes a smallpox-like human disease with a certain mortality rate (Marennikova, Shchelkunov, 1998). CPV is also capable of causing a generalized infection with a fatal outcome in immunocompromised persons. However, both MPV and CPV differ from VAR by their incapability of causing large-scale human epidemic outbreaks (low contagiousness). However, these viruses might acquire such properties in nature through evolutionary changes. Consequently, it is interesting to compare organizations of several orthopoxvirus genes to find out whether these genes evolve independently or according to a common pattern and to determine evolutionary relationships between poxviruses and the variation limits of their individual genes.
This work is both experimental and computer-based. The goal of this work was to sequence orthopoxvirus structural genes and those determining their virulence and to carry out their computer analysis. Orthopoxvirus A27L, A56R, B8R, and C11R genes (according to the nomenclature of vaccinia virus (VAC) strain Copenhagen) were analyzed. A27L gene encodes the 14K fusion protein of surface membrane of intracellular mature virus (IMV), providing formation of the envelope and release from the cell of extracellular enveloped viruses (EEV). This protein forms a stable complex with protein A17 (VAC), activates cell-mediated immune response in infected organism, and is essential for formation of EEV. A56R gene encodes hemagglutinin (HA), an envelope glycoprotein of EEV and membrane of infected cell. This protein inhibits fusion of infected cells and activates proteolytically the infectivity of virions. It is known that HA interacts with protein F13 (VAC). B8R gene encodes a homologue of γ-interferon receptor, secreted from the cell. C11R encodes the protein that is a secreted growth factor (VGF). It is known that the viral growth factor is not essential for virus replication. However, the experiments on its genetic inactivation have demonstrated an essential role of this factor in increasing the virulence and stimulating the cell proliferation at the site of primary infection. Growth factor plays an important role in the mechanism of virus penetration causing local hyperplasia and increasing the number of metabolically active cells, accessible for viral infection (Fenner et al., 1989; Shchelkunov, 1996).

Methods and Algorithms

Polymerase chain reaction was used to amplify in vitro the selected loci of viral genome containing target genes A27L, A56R, B8R, and C11R of the orthopoxviruses in question for further sequencing. PCR primers were selected so that the DNA fragments synthesized would contain the full-sized genes. The program Oligo 3.3 (Breslauer et al., 1986) was used to calculate the primers. Various strains of variola, monkeypox, cowpox, vaccinia, ectromelia, and camelpox were used in

the work. The DNA fragments obtained were cloned in *E. coli* cells within a plasmid vector for further sequencing using either Maxam–Gilbert technique or direct Sanger sequencing in an ABI PRISM 310 Genetic Analyzer automated device.

Results

Accomplishment of this work allowed us to accumulate information on the structures of the genes in question and carry out their comprehensive computer analysis. We compared the nucleotide sequences of the corresponding genes and the deduced amino acid sequences they code for. The available published nucleotide sequences of other poxviruses were also involved in this analysis.

Using the program Clustal X (Thompson et al., 1997), multiple alignments of both the nucleotide and amino acids sequences were constructed and their phylogenetic relationships were analyzed. The method used is the NJ (Neighbour Joining) method of Saitou and Nei. Permutation analysis of the statistical significance of the dendrograms constructed was carried out. Bootstrap NJ tree used a method for deriving confidence values for the groupings in a tree. It involved making 1000 random samples of sites from the alignment.

Discussion

Analysis of the phylogenetic relationships of the strains studied according to the nucleotide sequences of the target genes in question has demonstrated that these viruses fall into natural groups complying with their species attribution. Note that rabbitpox virus falls into the group of vaccinia virus, which is in accordance with the data on rabbitpox virus origin. In the cases of molecular virulence factors, such as hemagglutinin, γ-interferon receptor, and viral growth factor, analysis of genetic distances has detected a higher degree of divergence within the group of variola virus. However, this difference was not revealed in the case of gene A27L, encoding a virion protein.

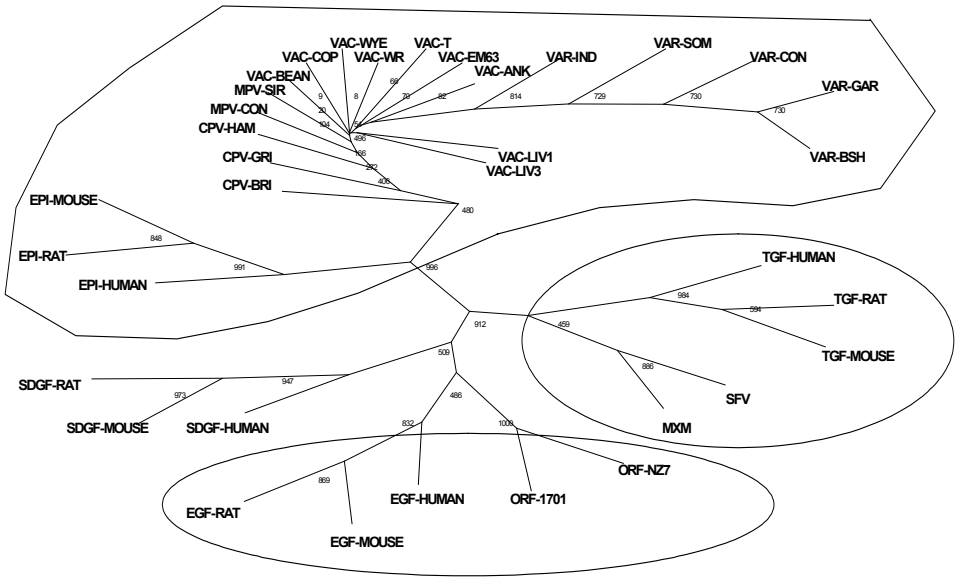


Fig. The unrooted tree constructed basing on amino acid sequences of C11R EGF-like region of poxviruses and cellular growth factors, such as EGF, SDGF, TGF, and EPI.

Differences in the phylogenetic relationships found between the same orthopoxviruses while analyzing individual genes is an important result of this study. Presumably, this indicates that individual orthopoxvirus species are capable of exchanging both gene fragments and larger regions of their genomes. The EGF-like region of C11R gene, encoding viral growth factor, is of the highest interest while analyzing this gene (Fig.). Its analysis has demonstrated that the variola viruses form a separate group on the dendrogram. Comparison of the orthopoxvirus growth factors with the growth factors of leporipoxviruses and orf viruses, heparin-binding epidermal growth factors (EGF), fibroblast- and astrocyte-derived growth factors (SDGF), growth factors expressed at early ontogenetic stages (epiregulins; EPI), and transforming growth factors (TGF) has demonstrated that the viral factors are most close to epiregulins. Leporipoxviruses, whose growth factors are close to transforming growth factors, form a separate group, whereas growth factors of orf viruses cluster with epidermal growth factors. This may indicate that different poxvirus genera had received different genes of the superfamily of cellular growth factors during their evolution.

Acknowledgements

Authors are sincerely grateful to J.J.Esposito, H.Meyer, S.S.Marennikova, and V.S.Petrov for the DNA preparations they had kindly provided. The work was partially funded by ISTC (grants № 884-2p and 1987) and Russian Foundation for Basic Research (grant № 00-04-49558).

References

1. Marennikova S.S., Shchelkunov S.N. (1998) Orthopoxviruses Pathogenic to Humans. M: KMK Scientific Press Ltd.
2. Shchelkunov S.N. (1996) The orthopoxvirus genome (a review). Mol. Biol. (Mosk.). 30, 5–32.
3. Fenner F., Wittek R., Dumbell K.R. (1989) The Orthopoxviruses. San Diego: Acad. Press Inc.
4. Breslauer K.J., Frank R., Blocker H., Marky L.A. (1986) Predicting DNA duplex stability from the base sequence. Proc. Natl Acad. Sci. USA. 83, 3746-3750.
5. Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acids Res. 24, 4876-4882.

COMPARATIVE STUDY OF THE ORTHOPOXVIRUS GENES B19R AND B29R

**** Mikheev M.V., Feshchenko M.V., Shchelkunov S.N.***

State Research Center of Virology and Biotechnology Vector, Koltsovo, Novosibirsk region, 633159 Russia,
e-mail: mihmv@vector.nsc.ru
*Corresponding author

Key words*: orthopoxviruses, phylogenetic analysis*

Resume

Motivation: Study of the structural organization of orthopoxvirus genome allows the genes controlling manifestation of pathogenicity, determining their host ranges and variation limits of individual genes to be discovered. In turn, these data will assist in formulating a model of the molecular evolution of orthopoxviruses.
Results: The sequencing and computer analysis of B19R and B29R genes of a large set of orthopoxvirus strains have demonstrated that despite certain differences observed, these sequences had evolved from a single ancestor. In the dendrograms constructed, the majority of orthopoxviruses form the groups coinciding with their species attribution except for cowpox viruses, falling into several rather distant subgroups. Comparison of the two phylogenetic trees makes evident a considerable distinction between them. In terms of genetic distances, B29R displays a considerably higher variation compared with B19R.

Introduction

The natural circulation of variola virus (VAR), causing numerous fatal outcomes during its epidemics, was halted in 1977 thanks to the Global Program of Smallpox Eradication. However, other viruses belonging to the genus Orthopoxvirus. family Poxviridae—monkeypox virus (MPV), cowpox virus (CPV), and buffalopox virus (BP) as well as ectromelia (mousepox; ECT) and camelpox (CML) viruses, nonpathogenic to humans—still continue persisting in nature. As vaccination against smallpox was ceased, the majority of present population lacks any immunity to orthopoxvirus diseases. The epidemic of human monkeypox in Zaire, continuing since 1996, is a confirmation of this fact. It is also known that CPV is capable of causing a generalized infection with fatal outcome in immunocompromised individuals.
The genome of orthopoxviruses encodes about 200 proteins; a part of these genes is responsible for virus replication, while other genes control manifestation of pathogenicity and determine the host range. The goal of this work was to sequence two orthopoxvirus genes—B19R and B29R (according to the nomenclature of vaccinia virus (VAC) strain Copenhagen), whose products are orthopoxvirus molecular virulence factors—and carry out their computer analysis. B19R encodes the protein binding to α/β-interferon and responsible for inhibition of the host protective reactions. B29R encodes the protein binding to numerous CC chemokines and inhibiting the development of host inflammatory and immune reactions to the infection. The protein B29R is produced in both secreted form and bound to virion surface. The protein B19R is produced in a secreted form.

Implementation

The genes selected were sequenced according to Sanger using an ABI PRISM 310 Genetic Analyzer automated device. Six oligonucleotide primers, calculated by the program Oligo 6, were used for sequencing B19R gene; four primers, for B29R. At the first stage, DNA fragments were amplified using the external primers followed by sequencing using the same primers and several internal primers. The data obtained were analyzed using the program Sequincher. DNAs of various orthopoxviruses—VAR, MPV, CPV, BP, VAC, ECT, CML, and rabbitpox virus (RPV), both sequenced by the authors and obtained from other laboratories—were used in the work. The program Mega, release 2.1 (Kumar et al., 2001) was used to construct dendrograms and calculate genetic distances.

Results and Discussion

Overall, B19R nucleotide sequences of 53 orthopoxvirus strains were determined; B29R sequences, of 58 strains. The program Clustal X, release 1.18 was used to align the sequences obtained and the relevant published sequences of other orthopoxvirus strains. In the phylogenetic trees and dendrograms constructed (Figs. 1, 2), the orthopoxviruses clustered according to their species attribution. Moreover, the genetic distances within these groups are small, except for the group of cowpox virus, and fall into the range of 0 to 0.006 in the case of B19R; of 0.001 to 0.004, in the case of B29R. Cowpox virus displays the most pronounced distinctions between its strains. These viruses form no separate group on the dendrograms: in the case of B28R, they form two subgroups with a distance of D = 0.082 between them; in the case of B19R, into three subgroups plus an additional separate strain with a distance between each amounting to 0.069.

Interestingly, all the species groups have their analogues among the cowpox virus subgroups. These results may be a confirmation of the hypothesis on evolution of all the modern orthopoxvirus species from cowpox or a cowpox-like virus. Comparison of the dendrograms (Figs. 1, 2) demonstrated that they were essentially different. For example, CPV strain GRI-90 in the B29R dendrogram appears most close to VAC and falls into the group with other CPV, whereas in B19R dendrogram, it is closest to ECT and is dissimilar to the other CPV strains. In addition, some other distinctions are more essential. For example, the genetic distance between the VAC and VAR groups calculated according to the B29R amounts to 0.03; VAC and MPV, 0.08; whereas the distance between VAC and VAR calculated according to B19R equaled 0.08; VAC and MPV, 0.045. Consequently, this suggests that individual genes of orthopoxviruses have evolved in parallel and independently.

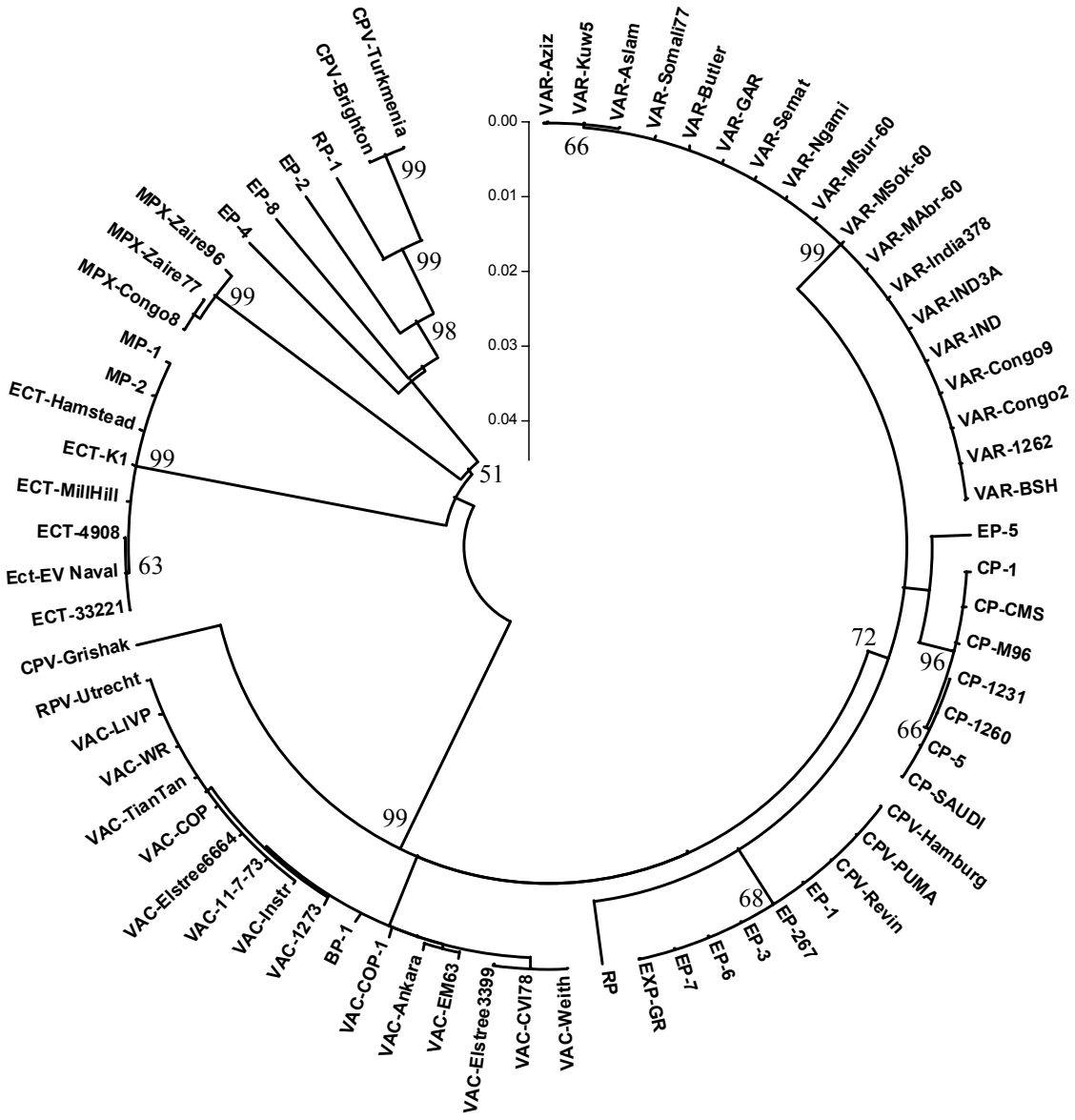


Fig. 1. Dendrogram constructed according to the sequence of B29R coding region using Minimum Evolution. Figures show results of realization bootstrap the analysis received at the analysis of 1000 trees. Values exceeding 50% are given only.

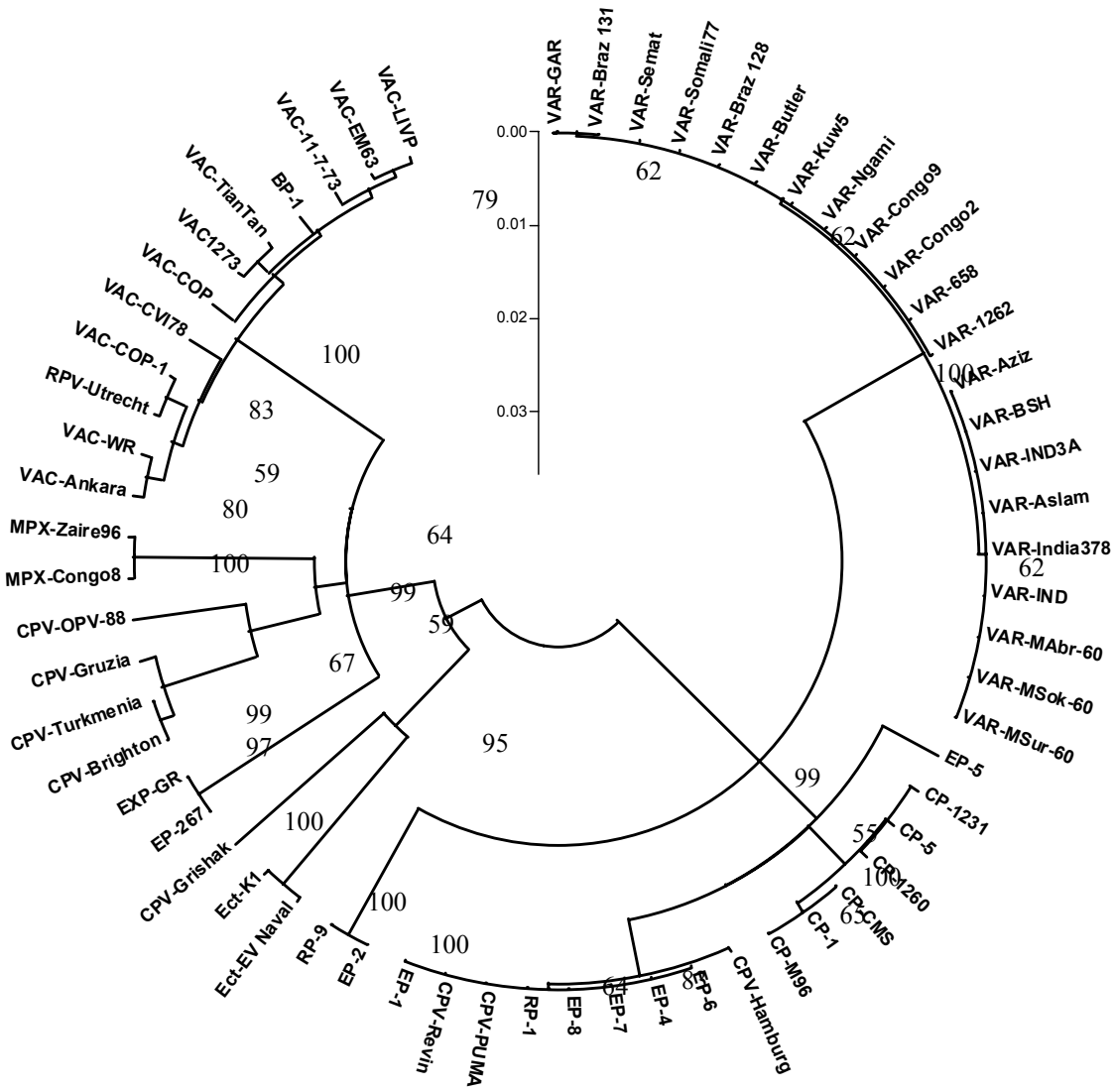


Fig. 2. Dendrogram constructed according to the sequence of B19R coding region using using Minimum Evolution. Figures show results of realization bootstrap the analysis received at the analysis of 1000 trees. Values exceeding 50% are given only.

Acknowledgements

Authors are sincerely grateful to J.J.Esposito, H.Meyer, S.S.Marennikova, and V.S.Petrov for the viral DNA preparations they had kindly provided. The work was partially funded by ISTC (grants № 884-2p, 1516, and 1987) and Russian Foundation for Basic Research (grant № 00-04-49558).

References

1. Kumar S., Tamura K., Jakobsen I.B., Nei M. (2001) MEGA2: Molecular Evolutionary Genetics Analysis software, Arizona State University, Tempe, Arizona, USA.

ANALYSIS OF BACTERIAL RM-SYSTEMS THROUGH GENOME-SCALE ANALYSIS AND RELATED TAXONOMY ISSUES

[1] Vandenbogaert M., [2] Makeev V.

[1] INRIA Rocquencourt - LaBRI Bordeaux I, Domaine de Voluceau, Le Chesnay 78153, France
[2] State Scientific Centre "GosNIIGenetika", Moscow, 3150501, Russia
e-mail: Mathias.Vandenbogaert@inria.fr,Makeev@imb.ac.ru

Key words: *restriction / modification systems, genome statistics, word usage biases, bacterial genomes, taxonomic inference*

Resume

Motivation: Recognition sites for type II restriction and modification enzymes in genomes of most bacteria are recognized as semi-palindromic motifs and are avoided at a significant degree. The key idea of contrast word analysis with respect to RMS recognition sites, is that under-represented words are likely to be selected against. Starting from contrast words corresponding to RMS recognition sites in specific clades, the specificity of unknown RMS can be highlighted. Eventually, this motivates the assessment of horizontal transferring events of RMS through the analysis of word usage biases in specific genomic regions. At last, the acquired observations can help to infer phylogenic relationships among bacteria.
Results: A probabilistic model is built on a first-order Markovian chain. Statistics on the *k*-neighborhood of a word are held to assess the biological significance of a genomic motif. Efficient word counting procedures are coupled to extremal statistics for assessing the significance of individual words in large sequences. A comparison of avoided palindromes in taxonomically related bacteria shows a pattern of relatedness of their RM-systems. To strengthen this analysis, the protein sequences of all type II RM-systems known in Rebase have been Blasted against the nr-Genbank database. The combination of these analyses has revealed some interesting examples of possible horizontal transfer events of RM-systems.

Introduction

Recognition sites for type II RM enzymes in genomes of several bacteria are semi-palindromes that are avoided at a significant degree, relating the avoidance of those short oligonucleotide words to RM systems (Panina et al., 2000). In the context of RM systems, distinct words that are recognized are selected against, creating so-called “contrast” words (Gelfand, Koonin, 1997). Statistically speaking, these contrast words appear to be avoided. From an evolutionary point of view it is argued that this is due to an occasional failure of methylation systems (Rocha et al., 1998), so that these bacteria that show a more stringent word usage bias do cope with less selective pressure than those that have words of a more random composition. Altogether, the correlation between contrasting individual 6-palindromes in specific taxa and the presence of closely related endonuclease and/or methylase enzymatic systems has not been systematically traced. Statistical methods have been widely used in determining the level of over- or under-representation of contrast words (Burge et al., 1992; Schbath et al., 1995; Gelfand, Koonin, 1997; Robin, Daudin, 1999; Beaudoing et al., 2000; Régnier et al., 2000; Denise et al., 2001) in the field of genomics. In this respect, we are focusing on pattern matching and analysis methods applied to semi-palindromes (that is: with well determined mismatches) in bacterial RM-systems.

Mathematical and Statistical Tools

Several combinatorial methods have already been used in terms of word counting (Nicodème et al., 1999). The mathematical formalism and computer science related aims lie within the context of word counting in large sequences and the assessment of their significance through extremal statistics. Extremal statistics have been used for scanning genomic sequences for signals that appear to be hidden by their neighbors (Denise et al., 2001). Fast approximating formulas have been elaborated, and have been implemented in the *QuickScore* library that will contain all necessary procedures that are used to this end.

Experimental Results

We model genomic sequences through a 1st order Markov chain. Using the word-counting procedures in the *QuickScore* library, statistics are held on plain hexamers and hexamers with specified errors. *Z*-scores are computed using dinucleotide frequencies for the calculation of the expectation, together with the poissonian approximation formula according to (Régnier et al., 2000).

Discussion

In the works of (Gelfand, Koonin, 1997; Panina et al., 2000), *Z*-scores have been computed for all words of length 6 in a number of bacterial genomes, establishing the correlation between the degree of avoidance of the words and most palindromes. Here, in extension of these experiments, a similar relation is set up between *Z*-scores for *approximate* words and palindromes. A comparison of the avoided palindromes in taxonomically related bacteria, shows a pattern of relatedness of their RM-systems, among other possible evolutionary events, that very likely took place through a horizontal transfer mechanism. We illustrate this with a clear-cut sample among various examples: the comparison of 5 species of the Enterobacteriaceae group, *Escherichia coli*, *Salmonella typhimurium* and *Salmonella typhi*, *Yersinia pestis* and *Buchnera* sp. As a matter of exception in this series, the evolution of the aphid endosymbiont *Buchnera*, during its adaptation to intracellular life, involved a massive reduction in its genome. In short, genome evolution of such symbiotic and parasitic bacteria results in both convergent and divergent changes, as can be highlighted by the presence of pseudogenes in genome sequences of the symbiotic bacteria *Buchnera aphidicola*, and parasitic bacteria. Convergent genome characteristics include reduction in genome sizes and lowered GC content values, which is exemplified by recent gene inactivation events and offers clues to the process of genome deterioration and host-cell adaptation (Silva et al., 2001). This can be true for processes inactivating RM-systems in *Buchnera*, since this species, while host-protected from bacteriophages, has lost the need for protecting RM-systems: except for some HemK-like methylases (tagged "other methylases", no known recognition sequences) no RMS is known for *Buchnera* sp. Thus, due to its endosymbiontic existence that preserves the organism from lethal external parasitic RM-systems and hence evolutionary pressure, the *Buchnera* sp. genome had to be left out. We observe that as we are dealing with statistics that are admitting some error in the words (while respecting their palindromic symmetry), the list of avoided words sorted on *Z*-score shows similarity in the list-tops (cfr. Table*)*, through comparison of both 3 genomes, *E. coli*, *S. typhimurium* and *Y. pestis.* The preservation of the rankings of the words, can be explained for organisms that are closely related by taxonomic branchings, by overall gene similarity, and hence avoidance of the same words, due to the presence of related RM-systems.

Table. Comparison of 1-neighbour hexamers in 3 Enterobacteriaceae genomes, with the 10 most avoided words of *E. coli* taken as reference (*limited dataset in this abstract)*.

E. coli					*S. typhim.*					*Y. pestis*				
1	GCGCGC	2475	BsePI	-74.005	5	CTGCAG	1031	PstI	-76.992	1	GCGCGC	1026	BsePI	-58.8668
2	CAGCTG	1778	PvuII	-68.1495	1	GCGCGC	5034	BsePI	-75.2475	5	CTGCAG	733	PstI	-57.2816
3	GGGCCC	68	ApaI	-65.1429	2	CAGCTG	802	PvuII	-67.9479	7	CCATGG	1361	NcoI	-55.222
4	CGCGCG	2127	(null)	-64.2957	9	CCTAGG	13	AvrII	-62.8786	15	GTGCAC	522	ApaLI	-51.2357
5	CTGCAG	958	PstI	-59.5874	7	CCATGG	645	NcoI	-60.6337	9	CCTAGG	122	AvrII	-47.7677
6	AGGCCT	605	StuI	-58.8212	20	CGATCG	1808	PvuI	-57.0352	2	CAGCTG	450	PvuII	-46.2519
7	CCATGG	612	NcoI	-58.4821	18	TGGCCA	1129	BalI	-52.0238		CGCGCG	697	(null)	-46.0037
8	GCATGC	588	SphI	-56.3641	25	GTATAC	518	SnaI	-48.4353	26	CGGCCG	538	XmaIII	-45.8447
9	CCTAGG	16	AvrII	-53.9918	6	AGGCCT	759	StuI	-47.8875	28	CCGCGG	679	SacII	-42.2405
10	GGATCC	495	BamHI	-53.0896	11	CCCGGG	573	SmaI	-47.4106	13	GCCGGC	516	NaeI	-42.147

For strengthening this fact, homologues of different restriction endonuclease and methylase proteic sequences appear to be encoded in strains belonging to a closely neighboring taxonomic branch. We exemplify with *Salmonella* sp.:

- the coding sequence of a *S. typhimurium* adenine methylase (Rebase M.StyDam), yields sequences of *E. coli*, *Y. pseudotuberculosis* and *Y. pestis*, as homologues in the Enterobacteriaceae group, and with a lower Blast-hitting score with *V. cholerae* (Vibrionaceae), *A. actinomycetemcomitans* and *P. multocida* in other Proteobacteria sp.,
- a HemK-like methylase (Rebase M.StyLHemKP; no known recognition sequence) yields sequences of *E. coli*, *Y. pestis*, *Buchnera* sp. within the Enterobacteriaceae group, and with a lower Blast-hit for Pasteurellales sp., and for Vibrionaceae sp.,
- StyLTI, a type III restriction enzyme, shows good homology with several Proteobacteria: *N. meningitidis* (β), *H. pylori* (δ/ε), *P. multocida* (γ) and *M. catarrhalis* (γ). Surprisingly, close homologues are found in *B. cereus* and *S. epidermidis*, although they are part of a divergent group: the Bacillus/Clostridium group, probably due to a recent horizontal transfer, as can be inferred from their sequence homology, codon usage bias and GC content difference. Evidence for this comes from the fact that RMS are often linked with mobile genetic elements. Therefore, these genes can cause genome rearrangements. They have been assigned a selfish behavior as they compete with each other for recognition sequences in post-segregational killing and super-infection exclusion (Kobayashi, 2001), emphasizing their preservation through lateral transfer.

Conclusion and Perspectives

This study extends previous studies undertaken at establishing phylogenetic relationships, starting from contrast words, with respect to RM-systems present in specific clades. These analyses can highlight among others, the lateral transfer events of RMS in bacteria, by studying positional word usage biases. The model used in this study is able to reflect related RMS in closely related taxa, and can reveal events that are classified as horizontal transfers. Still, this model is to be refined, with respect to the number of words contained in the set characterizing the RMS recognition sequences (cfr. IUPAC ambiguity codes s.s.). This, supplemented with information about divergent evolutionary rates among RMS families of proteins, will give a better and more refined understanding of genome organization, interaction and dynamics of evolution mechanisms through mobile elements, like transposons and plasmids.

Acknowledgements

We would like to thank Mireille Régnier and Mikhail Gelfand for useful discussions and support about both computational and phylogeny-related issues. This study was partially supported with the grant of the French-Russian Lyapunov Institute.

References

1. Beaudoing E., Freier S., Wyatt J., Claverie J., Gautheret D. (2000). Patterns of Variant Polyadenylation Signal Usage in Human Genes. Genome Res. 10, 1001-1010.
2. Burge C., Campbell A., Karlin S. (1992). Over- and underrepresentation of short oligonucleotides in DNA sequences. Proc. Natl Acad. Sci. (89), 1358-1362.
3. Denise A., Régnier M., Vandenbogaert M. (2001). Assessing statistical significance of overrepresented oligonucleotides. Proc. First Intern. Workshop on Algorithms in Bioinformatics, Aarhus, Denmark, August 2001; INRIA research report 4132, 85-97.
4. Gelfand M., Koonin E. (1997). Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. Nucl. Acids Res. 25(12), 2430-2439.
5. Kobayashi I. (2001). Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution. Nucl. Acids Res. 29(18), 3742-3756.
6. Nicodème P., Salvy B., Flajolet P. (1999). Motif statistics. European Sym. on Algorithms-ESA99, LNCS 1643, 194-211.
7. Panina E., Mironov A., Gelfand, M. (2000). Statistical analysis of complete bacterial genomes: Avoidance of palindromes and RM systems. Mol. Biol. 34(2), 215-221.
8. Régnier M., Lifanov A., Makeev V. (2000). Three variations on word counting. Proc. German Conf. on Bioinformatics, Heidelberg; submitted to BioInformatics, 75-82.
9. Robin S., Daudin J.J. (1999). Exact distribution of word occurrences in a random sequence of letters. J. Appl. Prob. 36(1), 179-193.
10. Rocha E., Viari A., Danchin A. (1998). Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. Nucl. Acids Res. 26(12), 2971-2980.
11. Schbath S., Prum B., de Turckheim E. (1995). Exceptional motifs in different markov chain models for a statistical analysis of DNA sequences. J. Comp. Biol. (2), 417-437.
12. Silva F., Latorre A., Moya A. (2001). Genome size reduction through multiple events of gene disintegration in Buchnera APS. Trends Genet. 17(11), 615-618.

FNR/DNR/ANR-REGULON IN GAMMA-PROTEOBACTERIA

Gerasimova A.V. [1*]***, Rodionov D.A.*** [1]***, Mironov A.A.*** [2]***, Gelfand M.S.*** [1,2]

[1] State Scientific Center GosNIIGenetika, Moscow, 113545, Russia
[2] IntegratedGenomics-Moscow, POBox 348, Moscow, 117333, Russia
[7]-(095)-3150156
e-mail: a_gerasimova@yahoo.com
[*]Corresponding author

Key words: *FNR, ANR, DNR, computer analysis, aerobic-anaerobic regulation*

Resume

Motivation: Comparative approach to computer analysis of regulatory signals allows one to predict new signals in bacterial genomes with high accuracy. A prediction is reliable whenever candidate signals are consistently observed in several related genomes.
Results: We describe the FNR-regulon of the *E. coli*, *Haemophilus influenzae*, *Vibrio cholerae, Salmonella typhi, Klebsiella pneumoniae, Yersinia pestis, Pasteurella multocida,* and *Actinobacillus actinomycetemcomitans* genomes and ANR/DNR regulon in the *Pseudomonas aueruginosa,*
Pseudomonas fluorescens, Pseudomonas putida, Pseudomonas syringae, Pseudomonas stutzer, and *Shewanela putrefaciens* genomes.

Introduction

FNR is a cytoplasmic O_2-responsive regulator consisting of two domains, sensor and DNA-binding regulator. It activates expression of genes that are required for anaerobic respiration and related pathways in gamma-proteobacteria.
FNR activates expression of several anaerobic enzymes, in particular, nitrate and nitrite reductases (anaerobic respiration) and pyruvate formate-lyase (anaerobic fermentation). Besides, FNR represses several genes encoding aerobic enzymes, such as cytochrome d ubiquinol oxidase and NADH dehydrogenase.
In *Escherichia coli*, expression of more than 120 genes that are included in the FNR modulon depends on alternation of the aerobic and anaerobic growth [1, 2].
The ortholog of FNR in *Pseudomonas aeruginosa* is ANR. This regulatory protein is required for the anaerobic growth of *Pseudomonas aeruginosa*. The sequences similar to the consensus FNR-binding motif (TTGAT….ATCAA) were found in the promoter regions of several genes for anaerobic metabolism of *Pseudomonas aeruginosa,* such as arginine deiminase pathway enzymes *(arcDABC),* nitrite reductase *(nirS*), nitric oxide reductase *(norCB),* and azurin *(azu)*.
ANR was experimentally shown to be necessary for denitrification, arginine deiminase activity and cyanide production of *P. aeruginosa* [3]. Another CRP/FNR-related regulator, DNR is essential for denitrification, ANR and DNR have similar binding signals, and it is impossible to determine what regulator would bind a candidate site by purely computational methods.
Besides, we assume that in some cases both regulators can bind to the same site.

Methods and Algorithms

Application of the comparative approach to the analysis of regulatory signals allows one to reliably predict new sites in bacterial genomes. Observation of candidate sites upstream of orthologous genes in several related genomes makes a prediction more significant. Here we use the comparative approach for the analysis of the FNR/ANR/DNR regulons of gamma-proteobacteria. Bacterial genomes were analyzed using the software package Genome Explorer [4].

Results and Discussion

We describe the FNR-regulon of the E. coli, Haemophilus influenzae, Vibrio cholerae, Salmonella typhi, Klebsiella pneumoniae, Yersinia pestis, Pasteurella multocida, and Actinobacillus actinomycetemcomitans genomes. The core of the regulon seems to be well conserved. Several new members were found in the FNR-regulon of Escherichia coli [5].
The candidate FNR binding sites were found upstream of twelve genes of E.coli that were known to be regulated by FNR. Fifteen new operons were predicted to be potential members of the FNR-regulon of E. coli, FNR-regulons in the other genomes were described for the first time.
In particular, the comparative analysis of Pasteurellaceae (H. influenzae, P. multocida, A. Actinomycetemcomitans) lead to identification of 26 candidate FNR-regulon genes in P. multocida.
We also described the members of ANR/DNR regulon in the Pseudomonas aueruginosa,
Pseudomonas fluorescens, Pseudomonas putida, Pseudomonas syringae, Pseudomonas stutzer, and Shewanela putrefaciens genomes. In Pseudomonas aueruginosa, the regulon contains about 30 candidate members.

Generally, the FNR regulons in Enterobacteriaceae, Vibrionaceae, and Pasteurellaceae are similar, and differ from the ANR/DNR regulon of Pseudomonads. However, there still exist genes that are members of these regulons in all studied genomes.

The FNR regulons of enterics/vibrio have a lot of common members and differ from the ANR regulons in pseudomonads. However, some genes belong to the FNR/ANR regulons in all genomes.

This study is the first attempt to describe global regulons in a large and diverse taxonomic group. Its results provide data to analysis of evolution of regulatory interactions in bacterial genomes. Additionally, new regulatory sites were predicted and regulons of several less studied genomes were described.

Acknowledgements

This study was partially supported by RFBR, INTAS and HHMI.

References

1. Bauer C.E., Elsen S., Bird T.H. (1999) Mechanisms for redox control of gene expression. Ann. Rev. Microbiol. 53:495-523.
2. Lynch A.S., Lin E.C.C. (1996) *Escherichia coli* and *Salmonella:* Cellular and molecular biology. Eds. Frederich C. Neindhard. Washington DC: ASM Pess, 1526-1538.
3. Hasegawa N., Arai H., Igarashi Y. (1998) Activation of a consensus FNR-dependent promoter by DNR of Pseudomonas aueruginosa in response to nitrite. FEMS Microbiol Lett. 166(2) :213-217.
4. Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000) Software for Analysis of Bacterial Genomes. Mol. Biol. 34:222-231.
5. Gerasimova A.V., Rodionov D.A., Mironov A.A., Gelfand M.S. (2001) Computer Analysis of Regulatory signals in Bacterial Genomes. Fnr-binding sites. Mol. Biol. 35:1001-1009.

SEARCH FOR REGULATORY SIGNALS IN GROUPS OF ORTHOLOGOUS GENES OF GAMMA – PROTEOBACTERIA

Danilova L.V. [1]***, Gelfand M.S.*** [2]

[1] Institute of information transmission problems, Moscow, Russia, e-mail: dlv2k@mail.ru
[2] Integrated Genomics-Moscow, Russia, e-mail: gelgand@integratedgenomics.ru

Key words: *gamma - proteobacteria, Orthologous genes, regulatory signals*

Introduction

Recognition of common regulatory signals in sets of DNA sequence fragments is an old and still actual problem of computational molecular biology. There are different approaches to this problem. One of them is analysis of upstream regions of orthologous genes from related genomes [1, 2]. The underlying assumption is that there is a conserved signal for orthologous regulators. This is not always correct, but in a sufficient number of cases this assumption holds, making the comparative technique a promising approach. Here we apply a previously suggested algorithm to analysis of genomes from the *Escherichia coli* group in order to test its applicability to other, less studied taxonomic groups.

Materials and Methods

Complete genomes of gamma-proteobacteria Escherichia coli, Escherichia coli O157, Salmonella typhi, Salmonella typhimurium, Yersinia pestis, Vibrio cholerae, Haemophilus influenzae, Pasteurella multocida were considered.
A pair of genes from two genomes was considered to be orthologous if these two genes were the closest relatives of each other in these two genomes. Then, pairs of orthologs were merged into clusters using the single linkage algorithm. Transitivity was not required and small differences in the similarity level were ignored (thus one gene could have more than one ortholog in any given genome).
Upstream regions of length 200 bp were selected. No overlaps with other genes were allowed, so if the distance to the upstream gene was shorter than 200 bp, only the spacer was selected.
Closely similar fragments were filtered out, retaining *E. coli* fragments whenever possible. This allowed us to search for conserved regulatory signals without interference from insufficiently divergent sequences from closely related genomes (strains). The criterion of excessive similarity was matches in at least 35 out of any 40 consecutive positions.
After the filtration step, there were 1967 subsamples of at least three fragments. After that, 345 sequences of known regulatory sites of *E. coli* were taken from the dpinteract database [3] and matched to the samples. Both directions of DNA sequence fragments were considered. Total 311 sites were found in 239 sequences. Other sites were not found either because they were located outside of the selected fragments or because the gene had no orthologs. A known sites could be placed in several samples if it was located between divergently transcribed genes, since each known site was considered in both direct and complementary directions.
The program implementing the earlier proposed algorithm [4] was used. It accepts as input some sequences (here 3 through 40) of length from 40 up through 200 bp, and outputs a system of similar words in each sequence. The system quality is defined by optimization of the pairwise similarity of words. It also takes into account additional features of words, e.g. their palindromicity. Signals of length 15, 20, 22, and palindromic signals of length 15, 16 and 22 were considered.

Implementation and Results

The results are shown in the Table. Ninety nine out of 311 known sites were found (that is, coincided with predicted signals or were subwords of the signals). Other sites were not found either becase the signals were too weak to be identified or because orthologs lost the regulation.

Discussion

The known sites *E. coli* considered for Table only. Our samples consist of region of orthologous genes from related *E. coli* genomes. If we have found site some regulator in *E. coli* fragments that means that we have found sites orthologuos genes the same regulator.
This work showed that such approach is reasonable for study taxonomic groups. We plan to apply that approach and our algorithm for *Bacillus subtilts* group and alpha-proteobacteria

Table. Nubmer of known sites *E.coli* in samples and detected sites.

Regulator	Number of known sites	Number of known sites in samples	Direct site	Complementary site	Detected sites
arcA	14	9	2	7	9
argR	17	20	14	6	3

cpxR	12	6	4	2	2
crp	49	41	26	15	3
cspA	4	6	3	3	1
cynR	2	4	2	2	1
cytR	5	4	4	0	0
deoR	3	1	1	0	0
dnaA	8	4	3	1	1
fadR	7	9	7	2	1
farR	4	8	2	6	4
fnr	14	10	6	4	1
fruR	12	6	3	3	4
fur	9	9	6	3	4
galR	7	4	3	1	2
gcvA	4	1	1	0	1
glpR	13	14	9	5	4
hns	15	11	6	5	4
hu	3	1	1	0	0
iclR	2	1	0	1	1
lacI	3	0	0	0	0
lexA	19	17	13	4	9
malT	10	17	9	8	5
melR	2	4	2	2	0
metJ	15	20	13	7	11
metR	8	10	7	3	1
narL	11	7	4	3	1
narP	8	10	6	4	0
ntrC	5	4	3	1	1
ompR	9	7	5	2	2
pdhR	2	2	2	0	1
purR	22	15	12	3	8
rpoN	6	4	3	1	3
torR	4	12	8	4	6
tyrR	17	13	13	0	5
Total	345	311	203	108	99

Acknowledgements

This study was partially supported by grants from INTAS (99-1476), HHMI (55000309), LICR (CRDF RBO-1268) and RFBR (00-15-99363). We are grateful to P.Novichkov for the help with the data, and V.A.Lyubetsky, the scientific leader of L.V.Danilova, K.Yu.Gorbunov and A.A.Mironov for discussions.

References

1. McCue L.A., Thompson W., Carmack C.S., Ryan M.P., Liu J.S., Derbyshire V., Lawrence C.E. (2001) Phylogenetic footprinting of transcription factor binding site in proteobacterial genomes. Nucl. Acids Res. 29, 3, 774-782.
2. Terai G., Takagi T., Nakai K. (2001) Prediction of co-regulated genes in Bacillus subtilts on the basis of upstream elements conserved across three closely related species. Genome Biol. 2, 11.
3. Robison K., McGuire A.M., Church G.M. (1998) A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome. J. Mol. Biol. 284, 241-254.
4. Danilova L.V., Gorbunov K.Yu., Gelfand M.S., Lyubetsky V.A. (2001) Algorithm of regulatory signal recognition in DNA sequences. Mol. Biol. 35, 6, 987-995.

TRANSCRIPTIONAL REGULATION OF A NEW BACTERIOCIN-PRODUCING SYSTEM IN *STREPTOCOCCUS EQUI.*

**** Kotelnikova E.A., Gelfand M.S.***

GosNII "Genetika", Moscow, Russia
IntegratedGenomics-Moscow
*Corresponding author

Key words: *transcriptional regulation, bacteriocin, genomics, Streptococcus equi*

Resume

Motivation: Bacteriocins production in several Gram-positive bacteria is controlled by two-component systems with different binding sites of transcriptional regulators. Genomic analysis provides an opportunity to predict putative bacteriocin-producing systems in less studied organisms.
Results: Response regulator (RR) of a two-component system (TCS), which controls the production of the class II bacteriocins in Gram-positive bacteria, binds to direct repeats with the standard unit length and period. We have analyzed the genome locus of *S.equi*, homologous to known bacteriocin loci in several Gram-positive bacteria. The signal search in *S.equi* reveals putative RR-binding sites upstream of TCS genes in this locus. Taken together, these observations allow us to predict the system of bacteriocins production in *S.equi* and its regulation.

Introduction

It is known that most if not all bacteria are capable of producing a heterogeneous array of molecules that may be inhibitory either to themselves or to other bacteria. These molecules include bacteriocins and bacteriocin-like substances that are directly produced as ribosomally synthesized polypeptides or precursor polypeptides (Jack, Tagg, 1995; Guder et al., 2000; Nes, Holo, 2000; Kleerebezem et al., 1997). Recently this type of antimicrobal peptides (AMPs) attracted considerable interest because of its potential usage in food preservation as well as in medical applications. The bacteriocins from Gram-positive bacteria are commonly divided into three groups: class I, the lantibiotics (Jack, Tagg, 1995; Guder et al., 2000); class II, heat-stable small (<10 kD) unmodified bacteriocins (Nes, Holo, 2000), and class III, larger heat-labile bacteriocins.
Most bacteria are known to rely on quorum-sensing systems as a cue for bacteriocin production. The quorum sensing is regulation of gene expression in response to fluctuations in the cell-population density. Quorum sensing in Gram-positive bacteria is usually carried out by typical bacterial two-component regulatory systems, consisting of a membrane-bound histidine kinase and a response regulator. Post-translationally processed signal peptide is secreted by a dedicated ATP-binding-cassette exporter and triggers the effect of two-component system. These signal molecules interact with the sensor element of the histidine kinase. In response, the histidine kinase autophosphorylates a conserved histidine residue within its intracellular domain. Subsequently, the phosphate group is transferred to the response regulator. Then the response regulator undergoes a conformational change that enables the C-terminal domain to bind to the operator region of the respective gene. Consequently, transcription of the regulated genes is activated or repressed. In some cases the bacteriocins themselves serve as signals, thus autoregulating their own synthesis and functioning as "quorum sensing" molecules.
Several regulated promoters were mapped within gene clusters involved in the production of class II AMPs in various bacteria. Sequence alignment of sites in each genome revealed the presence of a direct repeat of 9 or 10 nucleotides that are separated by 12-14 nucleotides and located 2-9 bp upstream of the –35 region. These repeats represent the binding sites for the corresponding response regulator.
In this study we used comparative genomics to predict gene clusters involved in the production of class II AMPs in Gram-positive bacteria.

Methods and Algorithms

Similarity searches against the NCBI protein database were performed using the BLAST program (Altschul et al., 1990). SignalX and GenomeExplorer (Mironov et al., 2000) was used for identification of the response regulators binding sites and genome comparison. GenBank sequences of *Streptococcus pneumoniae*, *Streptococcus equi* and *Streptococcus thermophilus* were studied.

Implementation and Results

First, we used BLAST and GenomeExplorer to identify genes potentially significant for bacteriocin production. To do that, all coding regions homologous to known genes of signal peptides, transporters, two-component systems and immunity in bacteriocin clusters were analyzed. As a result, we have found two gene clusters in *Streptococcus equi* containing response regulators highly similar to *blpR* from *Streptococcus pneumoniae* (Zaizieu et al., 2000) and several positionally linked genes encoding, in particular, histidine kinase and transporters.

To find potential binding sites of response regulators in *S.equi* we collected upstream regions of known operons in bacteriocins clusters. Then regulatory repeats of known type (9-10 bp separated by 12-14 bp) were analyzed in these regions using SignalX. They were aligned and search profiles were constructed for each genome. These profiles were then used for identification of candidate RR-binding sites in selected DNA sequences. Several direct repeats were observed in the predicted bacteriocins locus of *S.equi*. Candidate binding sites were observed upstream genes of transporter, RR, and bacteriocin-like peptide. The training set of *S. pneumoniae* sites and the predicted binding sites from *S. equi* are shown in Table 1.

Table 1. Potential response regulators binding sites in *S. pneumonia* and *S. equi*.

Organism	Genes	Sequences
S. pneumoniae	*blpX*	**ATTCAAGATGT**ttcgatgaca**ATTCAAGATTT**
	blpU	**ATTCAAGACGT**ttcgatgcca**ATTCAAGATTT**
	blpI	**ATTCAAGACGT**ttcgatgaca**ATTCAAGATCT**
	blpM	**ATTCAAGACGT**ttcgatgact**ATTCAAAATCT**
	blpA	**ATTCAGGAAGT**tttaatgact**ATTCAAGATTT**
	blpT	**ATTCAAGACAT**ttcaatgaca**ATTAAAGATTT**
	blpL	**ATTCAAGAGGT**tttgatgacc**ATTTATGATTT**
S. equi	*blpA_1*	**ATTTAAGACGT**ttcaacgact**ATTCAAGACTA**
	blpA_2	**ATTCAAGACGT**ttcgacgaca**TTTTAAGACTT**
	blpM	**ATTCAAGACGT**ttcgacgaca**TTTTAAGACTT**
	blpR1	**ATTTAAGACAT**aatcagtacc**ATTTAAGATTT**
	transporter	**ATTCACGACAA**aataagaacc**ATTCAAGATTT**

Additional sequence similarity search using DNA-protein alignment revealed two open reading frames in the *S.equi* bacteriocin cluster. These ORFs are homologous to the genes *thmA* and *thmB* from *S.thermophilus* (Fig.) encoding two-peptide bacteriocin thermophilin 13. Furthermore, genes *blpM* and *blpN* encoding bacteriocin-like peptides in *Streptococcus pneumoniae* have strong sequence similarity to another region in *S.equi* bacteriocins locus, which was not annotated.

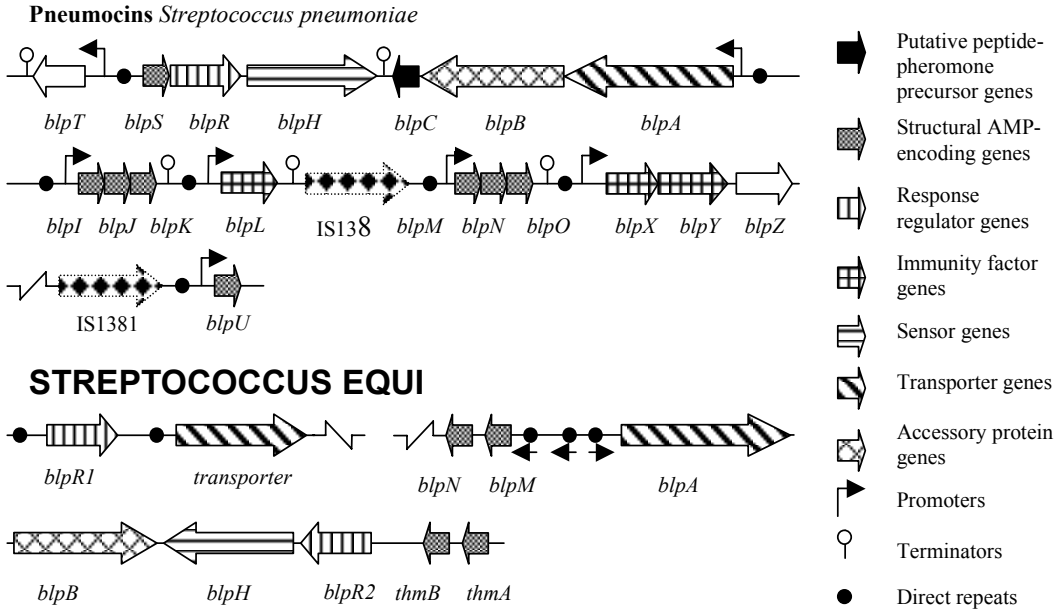


Fig. Genetic organisation of bacteriocins clusters in *S. pneumoniae* and *S. equi*.

Discussion

We have identified two gene clusters of *S. equi* which have strong similarity to known genetic determinants of bacteriocin production in Gram-positive bacteria. Both response regulators in *S. equi* are highly similar to RR BlpR from *S. pneumoniae*. Additionally, genes of bacteriocin-like proteins were mapped. This allows us to predict the bacteriocins production system in *S. equi*. To support this suggestion, the putative regulatory system in *S. equi* was analyzed. Site search revealed direct repeats similar to repeats from *S. pneumoniae* which are likely binding site of the corresponding RR, as they are located upstream of genes in the putative bacteriocin clusters.
It is common for bacteriocins loci in other bacteria to have RR, histidine kinasw and signal peptide in the same operon. In one *S. equi* locus only RR and transporter were found, and the other locus contains all necessary genes for the bacteriocins production except the signal peptide. However, the genome of *S. equi* is still incomplete and thus these specific features could in fact be an artifact caused by contig breaks.

Acknowledgements

The work was supported by the INTAS grant № 99-1476 and HHMI grant № 55000309. The authors are grateful to A.B.Rachmaninova and O.Laikova for useful discussions.

References

1. Altschul S.F., Gish W. et al. (1990) Basic local alignment tool. J. Mol. Biol. 215, 403-410.
2. Guder A., Wiedemann I., Sahl H.-G. (2000) Posttranslationally modified bacteriocins – the lantibiotics. Biopolymers (Peptide Science). 55, 62-73.
3. Jack R.W., Tagg J.R., Ray B. (1995) Bacteriocins of Gram-positive bacteria. Microbiological Reviews. 59, 171-200.
4. Kleerebezem M., Quadri L.E.N. et. al. (1997) Quorum sensing by peptide pheromones and two-component signal-transduction systems in Gram-positive bacteria. Mol. Microbiol. 24, 895-904.
5. Mironov A.A., Vinokurova N.P., Gel'fand M.S. (2000) Software for analyzing bacterial genomes. Mol. Biol. (Mosk). 34**,** 253-262.
6. Nes I.F., Holo H. (2000) Class II antimicrobial peptides from lactic acid bacteria. Biopolymers (Peptide Science). 55, 50-61.
7. Zaizieu A., Gardes C. et al. (2000) Microarray-based identification of a novel Streptococcus pneumoniae regulon controlled by an autoinduced peptide. J. of Bacteriology. 182**,** 4696-4703.

SYSTEMATIC PREDICTION OF REGULATORY INTERACTIONS IN THE LACI FAMILY OF TRANSCRIPTIONAL REGULATORS

Laikova O.N.

State Scientific Center GosNIIGenetika, Moscow, 113545, Russia, e-mail: laikova@mail.ru

Key words: *LacI family; transcription regulation; comparative approach*

Resume

Motivation: The challenge of elucidating the DNA-protein recognition mechanisms requires that a substantial number of regulatory proteins and their binding sites are known. With the massive sequencing of prokaryotic genomes, the number of putative transcriptional regulators, identified by similarity to known regulators, is constantly increasing. On the other hand, the number of experimentally confirmed binding sites is rather low.
Results: We considered the LacI family of regulators and utilized phylogenetic, positional and functional data in the context of genomic comparison in order to systematically predict candidate binding sites for the regulators in sequenced bacterial genomes. As a result, we have created an inventory of the LacI family, predicted a number of candidate binding sites, predicted several regulons *de novo*, and made some interesting observations regarding shifts of the regulators' specificity for binding sites and effectors.

Introduction

Transcriptional regulators containing the Helix-Turn-Helix (HTH) motif for recognition of DNA sites are divided into families on the base of amino acid sequence similarities. One of these families, the LacI family, was first described in 1992 (Weickert, Adhya, 1992). At that time the family included 17 full-length proteins and 2 partial sequences lacking the HTH motifs. The characteristic feature of the LacI family is significant similarity throughout the whole length of the sequences and high conservation of the N-terminal HTH motif. The C-terminal effector-binding domains of the regulators are homologous to some substrate-binding proteins of ABC-type transporters (RbsB, AraF, MglB).
Some presumptions can be drawn from the available experimental data. Firstly, the substantial fraction of the LacI family regulators are local regulators, e.g. LacI, RbsR, TreR, CscR of *Escherichia coli*. Secondly, because the regulators bind DNA in the dimeric form, the binding sites demonstrate two-fold symmetry, being either even or odd inverted repeats, although there are some deviations, e.g. the well-known CytR regulator. An additional helpful observation is that metabolic pathways whose structural genes are regulated by the LacI family proteins are often peripheral pathways for the utilization of sugars and their derivatives. We combined these presumptions with the comparative approach to the recognition of transcription regulatory sites. The approach is based on the assumption that when a regulator gene is conserved in several genomes, the sets of structural genes regulated by this regulator are conserved as well. Using the above ideas along with phylogenetic study of the LacI family, we have predicted new transcription regulation sites in already known regulons, as well as a number of regulons *de novo*, linking regulators to regulated genes and operators.

Methods and Algorithms

Published microbial genomes were downloaded from the EMBL/GenBank databases. Preliminary sequence data for bacterial genomes were obtained from servers of the DOE Joint Genome Institute (http://www.jgi.doe.gov/JGI_microbial/html/index.html), the Sanger Institute (http://www.sanger.ac.uk/Projects/Microbes/), the Institute for Genomic Research (http://www.tigr.org/), the Genome Sequencing Center at the Washington University (http://genome.wustl.edu/gsc/Projects/bacteria.shtml), and the University of Oklahoma's Advanced Center for Genome Technology (http://www.genome.ou.edu).
The initial sample of the LacI family members was derived from the SWISS-PROT and TrEMBL databanks using the SRS web server (http://srs.ebi.ac.uk/).
Similarity searches in the genomes were performed using the Smith-Waterman algorithm as implemented in the GenomeExplorer program (Mironov et al., 2000). The homology relationships of proteins were confirmed by the InterPro database (http://www.ebi.ac.uk/interpro/scan.html).
Phylogenetic trees were constructed with ClusalX (the neighbor joining method) and the ProML program of the PHYLIP package (the maximum likelihood method).
In order to construct positional nucleotide weight matrices (profiles) from a set of DNA fragments representing potential regulatory regions, an iterative procedure implemented in the SignalX program was performed (Gelfand et al., 2000). According to the procedure, weak palindromes are selected in each DNA fragment, then each palindrome is compared with all others, and the palindromes most similar to the initial one are used to make a profile. The equation (1) defines positional nucleotide weights in a profile:

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

where *N(b,k)* is the count of nucleotide *b* at position *k*. The site score is the sum of the respective positional nucleotide weights. The base of the logarithm was chosen such that the standard deviation of the site score distribution on random oligomers equals 1 (Mironov et al., 1999). With these profiles the set of palindromes is scanned again, and the procedure is iterated until convergence. The quality of a profile of the finally derived set of profiles is defined as its information content:

$$I = \sum_{k=1\ldots L} \sum_{i=A,T,G,C} f(i,k) \log(f(i,k)/0.25) \quad (2)$$

where *f(i,k)* is the frequency of nucleotide *i* at position *k* of the palindromes generating the profile, and *L* is the palindromes length. The best profile is used as a recognition rule for scanning the studied genomes.

Implementation and Results

The initial sample of the LacI family proteins extracted from the SWISS-PROT and TrEMBL (above 80 proteins) was used to find homologous sequences in the studied genomes. At present, we have considered 63 genomes containing putative LacI family regulators, as well as a number of DNA fragments extracted from EMBL/GenBank. As a result, there are more than 550 representatives of the LacI family. They are found only in Bacteria, but not Archaea or Eukaryota.

The phylogenetic tree constructions and pair-wise genome comparisons were used to assign approximately 400 proteins into about 100 orthologous groups containing from 2 through 29 members. The remaining proteins have no obvious orthologs in the available genomes. The orthologous groups containing only two proteins were not systematically considered.

The loci adjacent to the orthologous regulator genes were compared in several genomes to identify candidate regulated operons containing orthologous structural genes. The putative operons were selected, considering the direction of transcription and the lengths of intergenic regions (no more than 200 bp).

The recognition signals were identified by applying SignalX to samples of upstream (regulatory) regions. Finally, additional candidate regulatory sites were identified by scanning genomes with the constructed profiles. As a result, for about 250 regulators representing more than 50 orthologous groups, the regulons have been predicted or expanded. Now we maintain a collection of the LacI family profiles and operators, which are readily updated as new bacterial genomes become available.

Table. Distribution of the LacI family regulators in bacteria (abridged).

Genome	Status[1]	I	II	III
Caulobacter crescentus	CG	12	3	1
Rhodobacter sphaeroides	UG	6	6	6
Rhodobacter capsulatus	UG	4	4	4
Mesorhizobium loti	CG	20	13	9
Sinorhizobium meliloti	CG	29	19	13
Agrobacterium tumefaciens C58	CG	21	16	12
Brucella melitensis	CG	6	5	5
Bordetella parapertussis	UG	5	3	1
Ralstonia solanacearum	CG	6	6	5
Ralstonia metallidurans CH34	UG	2	2	1
Burkholderia pseudomallei	UG	6	5	4
Burkholderia fungorum LB400	UG	11	9	5
Pseudomonas aeruginosa	CG	4	4	4
Pseudomonas fluorescens Pf0-1	UG	5	5	5
Pseudomonas syringae pv. tomato DC3000	UG	6	6	6
Escherichia coli K-12	CG	14	14	13
Salmonella typhimurium	CG	14	14	10
Klebsiella pneumoniae MGH78578	UG	28	26	15
Yersinia pestis	CG	17	16	8
Yersinia enterocolitica	UG	20	20	10
Pectobacterium carotovorum atrosepticum	UG	20	19	9
Pasteurella multocida	CG	6	6	5
Actinobacillus actinomycetemcomitans	UG	4	4	3
Haemophilus influenzae	CG	3	3	3
Vibrio cholerae	CG	11	11	10
Vibrio fischeri	UG	9	9	6
Corynebacterium glutamicum ATCC 13032	CG	7	2	0
Corynebacterium diphtheriae	CG	2	2	0
Thermobifida fusca	UG	6	4	1
Streptomyces coelicolor	CG	35	11	1
Staphylococcus aureus N315	CG	4	4	3
Bacillus subtilis	CG	11	10	6
Bacillus halodurans	CG	14	11	6

Bacillus cereus ATCC 14579	CG	9	5	2

Continuation

Bacillus stearothermophilus	UG	10	9	4
Listeria monocytogenes	CG	11	8	3
Streptococcus pneumoniae TIGR4	CG	7	6	3
Streptococcus pyogenes M1	CG	6	6	2
Streptococcus mutans	CG	5	5	3
Lactococcus lactis	CG	5	3	3
Enterococcus faecium DO	UG	10	8	4
Enterococcus faecalis V583	UG	12	10	6
Lactobacillus gasseri	UG	5	5	3
Clostridium acetobutylicum	CG	7	5	2
Clostridium perfringens	CG	7	5	3
Clostridium difficile	UG	6	3	2
Thermotoga maritima	CG	5	5	0
Petrotoga miotherma	UG	7	6	1
Deinococcus radiodurans	CG	2	1	1
Thermus thermophilus	UG	2	1	1
Total		484	383	233

[1]CG and UG are for completely sequenced and unfinished genomes respectively. Columns I, II, and III show: total number of genes encoding LacI family regulators in a genome, number of the regulators which have orthologs, and number of the regulators whose binding sites are known or predicted in this work.

Discussion

The comparative approach to the prediction of transcription regulation requires that a regulator has orthologs in several genomes. When the regulator belongs to a large protein family, such as the LacI family, whose members share whole-length similarity and often are present in one genome as several paralogs, the pair-wise genome comparison does not allow to resolve the orthology relations. In such cases it is necessary to consider the entire family of regulators, and the phylogenetic tree construction is a helpful tool for finding most likely orthologous groups.

The LacI family members are in many cases regulators of peripheral metabolic pathways. Not surprisingly, the respective regulons are quite unstable, and genetic shifts, such as deletions and duplications, are common in these regulatory systems. Thus, in several cases, we have observed a regulator obviously belonging to a group of orthologous regulators for which the regulated genes were known in other genomes, but could not be found in the considered one. For example, the *YPO1642* gene of *Yersinia pestis* belongs to the CscR group of sucrose utilization regulators, whereas no orthologs of structural sucrose utilization genes can be found in the *Y. pestis* genome. The opposite situations, i.e. recent loss of a regulator and conservation of structural genes, were also observed.

In a sense, the peripheral pathways for utilization of various carbon and energy sources (feeders) are composed of three types of elements: regulators, transporters and enzymes, and these elements evolve rather independently. Thus, one can observe probable non-orthologous replacements of elements of each group. Non-orthologous regulators, which control the same metabolic pathway, can belong to the same protein family: e.g. at least four orthologous groups of LacI family regulators, quite distantly related to each other, include regulators of sucrose utilization in different bacteria; similarly, there seem to be at least six groups of LacI family regulators concerned with ribose utilization. Such regulons seem to be physiologically convergent, whereas true molecular effectors and DNA signals recognized by the regulators of different orthologous groups can diverge.

Although it does not seem possible to divide the LacI family into subfamilies "from the root", in some cases several orthologous groups might be gathered together into wider groups, whose phylogenetic history (duplications and divergence) might be reasonably guessed. Such groups offer interesting opportunities to study co-evolution of regulators, DNA signals recognized by the regulators, and the effector specificity of regulators.

Acknowledgements

This work was partially supported by the grants from INTAS (99-1476), HHMI (55000309), and RFBR (00-15-99362). I am grateful to M.S.Gelfand for influential discussions and inspiration.

References

1. Gelfand M.S., Koonin E.V., Mironov A.A. (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. Nucl. Acids Res. 28, 695-705.
2. Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucl. Acids Res. 27, 2981-2989.
3. Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000) Software for analysis of bacterial genomes. Mol. Biol. (Mosk). 34, 222-231.
4. Weickert M.J., Adhya S. (1992) A family of bacterial regulators homologous to Gal and Lac repressors. J. Biol. Chem. 267, 15869-15874.

PREDICTION OF NEW ENZYME INVOLVED IN PEPTIDOGLYCAN RECYCLING

[1]* Panina E.M., [2] Vassieva O., [1] Gelfand M.S., [2] Overbeek R.

[1] Integrated Genomics-Moscow, Institute of General Genetics, 119991, Moscow, Russia
[2] Integrated Genomics, Inc., 2201 W. Campbell Park Dr Chicago, IL 60612
e-mail: katya@ekpanina.mccme.ru
*Corresponding author

***Key words**: cell wall, murein, transcription regulation*

Resume

We demonstrate here a bioinformatics approach to prediction of a new enzyme involved in peptidoglycan recycling in bacteria. *YjjK* encoding this enzyme is frequently co-localized with the genes involved in murein turnover. Moreover, in the genome of *E. coli,* a common candidate regulatory sequence was found in the upstream region of *yjjK* as well as two other genes involved in cell wall biogenesis, namely *mltE* and *ampC*

Introduction

Turnover and recycling of the cell wall murein represent a major, though often non-essential, metabolic pathway of most bacteria. Degradation products of the peptidoglycan are formed during the enlargement of the murein sacculus as a consequence of a growth mechanism, which couples the controlled degradation of the cell wall polymer with the insertion of new material. Consequently, the recycling pathway is viewed as a possible signaling vehicle, informing the cell of the condition of the essential structure existing outside the cell itself (Park, 1995; 1996). Alginate production, bacterial encystment in *Azotobacter vinelandii* and induction of *Escherichia coli* β-lactamase genes were shown to be greatly influenced by the bacterial ability to recycle their cell wall (Nunez et al., 2000; Tuomanen, 1991; Tolg et al., 1993; Normak, 1995; Park, 1996; Dietz, 1997; Wiedemann et al., 1998).

β-Lactamase (AmpC; Edlund et al., 1979) induction and modulation of the composition of the cell wall share elements of a regulatory circuit that involves AmpD, cytosolic N-acetyl-anhydromuramyl-*L*-alanine amidase (Jacobs, 1995). Amidases were shown to act as powerful autolytic enzymes in the presence of antibiotics (Heidrich et al., 2001). Cell wall turnover products may relay the signal to AmpR, the ampC transcription activator (Lindberg et al., 1985; Lindquist et al., 1989) or act on AmpR indirectly through the AmpE member of AmpD/AmpE signal system (Park, 2001). There is also a possible connection between septation/division and induction of ampC β-lactamase promoted by ftsZ (Ottolenghi, Ayala, 1991).

The released cell-wall peptides are regulated by the Opp system in gram-negative bacteria and highly homologous Spo system in gram-positive (Goodell, Higgins, 1987; Perego et al., 1991). The MppA protein is responsible for tripeptide uptake (Li, Park, 1999) in gram-negative bacteria. In *E. coli,* the transmembrane protein AmpG (Lindquist et al., 1993) transports not only D-tripeptide but also D-pentapeptide into the cell (Park, 2001). Still, many aspects of this system should be yet discovered and understood.

Knowledge about the machinery performing regulation, turnover, and recycling of cell-wall components in bacteria seems to be of a major importance in designing inhibitors that could prevent the establishment of β-lactam resistance of bacteria possessing inducible β-lactamases. It also can help in developing new classes of antibiotics. We demonstrate here a bioinformatics approach to prediction of a new enzyme involved in the peptidoglycan recycling in bacteria. The *yjjK* gene, encoding this enzyme, is frequently co-localized with the genes involved in the murein turnover. Moreover, in the genome of *E. coli,* a common candidate regulatory sequence was found in the upstream region of *yjjK* and two genes involved in the cell wall biogenesis (Engel et al., 1992), namely *mltE* and *ampC*.

Materials and Methods

Genome sequences of analyzed species were extracted from the ERGO Database (http://wit.mcs.anl.gov/WIT2/). Profiles for signal recognition were constructed as described in (Gelfand, 1999). Positional nucleotide weights in these profiles are defined as

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

where $N(b,k)$ is the count of nucleotide b at position k. The score of the candidate site is calculated as the sum of the respective positional nucleotide weights:

$Z(b_1 \ldots b_L) = \Sigma_{k=1\ldots L}\, W(b_k,k)$, where k is the length of the site.

Genomic analyses (protein similarity searches using Smith–Waterman algorithm, analysis of orthology, and identification of candidate signals in the genome sequences) were done using GenomeExplorer (Mironov et al., 2000).

Results

The genome of *Escherichia coli* contains two paralogous genes, *yjjK* and *uup*, with an identity of 34%. Each of them encodes a protein previously identified as a putative ATP-binding component of a transport system. We have analyzed these proteins with PROSITE motif-search tool and TMPRED server. Each protein consists of two homologous parts; each part has a nucleotide-binding domain typical of ABC transporters, and, in the case of YjjK, a hydrophobic transmembrane alpha helix (predicted by TMPRED, Fig. 1a). The TMPRED output for Uup is less clear and predicts two transmembrane alpha helices in the N-terminal part of the protein (Fig. 1b). However, ABC transporters generally have at least four transmembrane alpha helices; thus, Uup and YjjK are more likely to be membrane-bound ATP-binding enzymes.

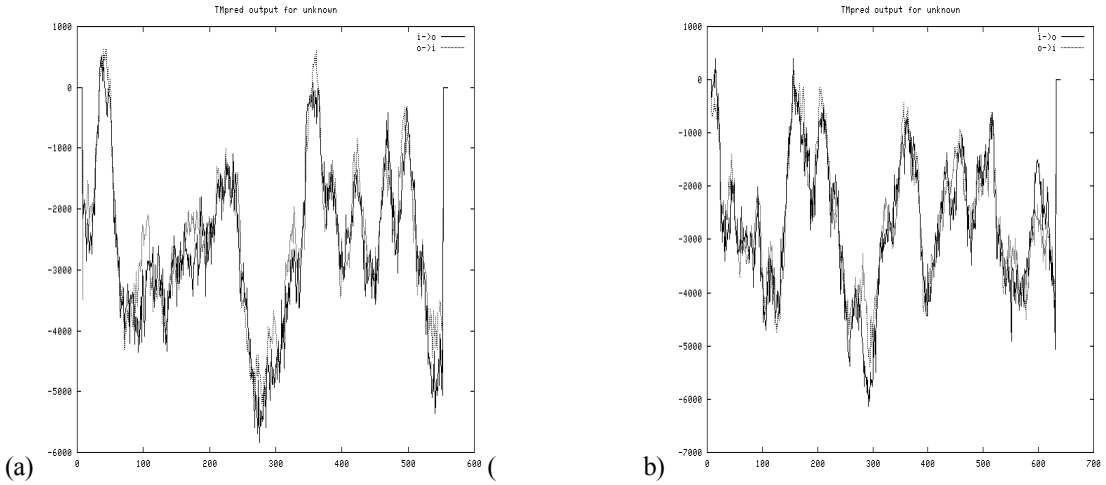


Fig. 1. TMPRED output for (**a**) yjjK and (**b**) uup proteins of *E. coli*.

We have found that *yjjK* and *uup* genes are frequently co-localized with the genes involved in murein turnover (Fig. 2). In particular, in the genomes of *Pasteurella multocida*, *Actinobacillus actinomycetemcomitans,* and *Haemophilus influenzae, uup* clusters positionally with the gene encoding a murein hydrolase exporter; in the genomes of *Pseudomonas aeruginosa* and *Pseudomonas fluorescens,* it is co-localized with a gene encoding soluble lytic murein transglycosylase. *yjjK* is adjacent to a soluble lytic murein transglycosylase gene in the genomes of *E. coli*, *Salmonella typhi,* and *Vibrio cholerae*. Moreover, in the genome of *Rhodopseudomonas palustris yjjK* is localized near a gene encoding a non-orthologous membrane-bound lytic murein transglycosylase.

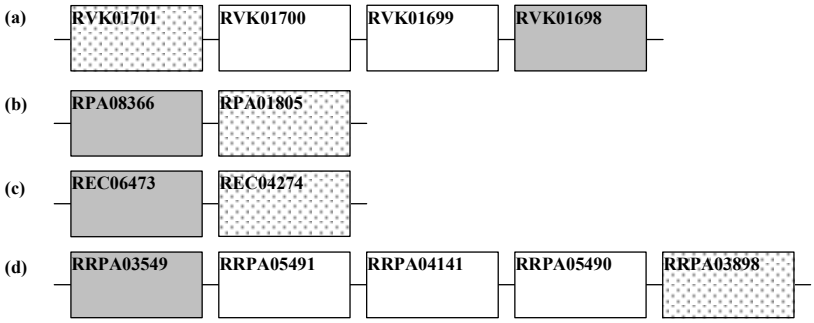


Fig. 2. Schematic representation of chromosomal loci containing homologs of *yjjK* and *uup* genes in (**a**) *P. multocida*, (**b**) *P. aeruginosa*, (**c**) *E. coli*, and (**d**) *R. palustris*. *yjjK* and *uup* orthologs are shown by filled arrows, genes functionally associated with murein are shown by hatched arrows. Gene identifications are given from ERGO database.

RVK01701 and *RPA01805* are orthologs of *uup*; *REC04274* and *RRPA03898* are orthologs of *yjjK*; *RVK01700* encodes an unknown hypothetical protein; *RVK01699* encodes deoxyguanosine triphosphate triphosphohydrolase; *RVK01698* encodes murein hydrolase exporter; *RPA01805* and *REC04274* encode soluble lytic murein transglycosylase; *RRPA05491* encodes an unknown hypothetical protein; *RRPA04141* encodes trans-aconitate methyltransferase; *RRPA05490* encodes DNA polymerase III, chi subunit; and *RRPA03898* encodes membrane-bound lytic murein transglycosylase.

In the genomes of *E. coli* and *S. typhimurium,* a common candidate regulatory sequence was found in the upstream region of *yjjK* as well as two other genes involved in cell wall biogenesis, namely *mltE* and *ampC* (Fig. 3). This sequence is a palindrome of length 20 with conserved 7-bp half-sites and a variable 6-bp spacer. This element is not conserved in other species.

```
yjjK     CTcaTTA - 6 - TAAaCAt
mltE     CTGtTTA - 6 - TAAcCcG
ampC     CcGGTTt - 6 - aAAcCAG
```

Fig. 3. Candidate regulatory sequence in the upstream regions of *yjjK, ampC,* and *mltE* genes of *E. coli.*

Being conserved among most prokaryotic as well as several eukaryotic species, both *yjjK* and *uup* genes are absent in the genomes of intracellular endosymbionts, such as *Mycoplasma genitalium* and *Mycoplasma pneumonia,* that lack the cell wall. Moreover, these genes are conserved in those eukaryotic species that posses the cell wall, e.g. *Arabidopsis* and *Drosophila*, and they are absent in mammals that has none.

Due to the co-localization of *yjjK* and *uup* genes with genes involved in murein recycling, their candidate co-regulation with cell wall genes, and finally, their absence in the genomes of species lacking the cell wall, but not in the species with the cell wall, we propose that YjjK and Uup are membrane-anchored ATP-binding proteins involved in the cell wall regeneration.

Acknowledgements

This study was partially supported by grants from INTAS (99-1476) and HHMI (55000309).

References

1. Dietz H., Pfeifle D., Wiedemann B. (1997). The signal molecule for beta-lactamase induction in *Enterobacter cloacae* is the anhydromuramyl-pentapeptide. Antimicrob. Agents Chemother. 41(10):2113-2120.
2. Edlund T., Grundstrom T., Normark S. (1979). Isolation and characterization of DNA repetitions carrying the chromosomal beta-lactamase gene of *Escherichia coli* K-12. Mol. Gen. Genet. 173(2):115-125.
3. Engel H., Smink A.J., van Wijngaarden L., Keck W. (1992). Murein-metabolizing enzymes from *Escherichia coli:* existence of a second lytic transglycosylase. J. Bacteriol. 174(20):6394-6403.
4. Gelfand M.S. (1999). Recognition of regulatory sits by genomic comparison. Res. Microbiol. 150:755-771.
5. Goodell E.W., Higgins C.F. (1987). Uptake of cell wall peptides by *Salmonella typhimurium* and *Escherichia coli.* J. Bacteriol. 169(8):3861-3865.
6. Li H., Park J.T. (1999). The periplasmic murein peptide-binding protein MppA is a negative regulator of multiple antibiotic resistance in *Escherichia coli.* J. Bacteriol. 181(16):4842-4847.
7. Heidrich C., Templin M.F., Ursinus A., Merdanovic M., Berger J., Schwarz H., de Pedro M.A., Holtje J.V. (2001). Involvement of N-acetylmuramyl-*L*-alanine amidases in cell separation and antibiotic-induced autolysis of *Escherichia coli.* Mol. Microbiol. 41(1):167-178.
8. Jacobs C., Huang L.J., Bartowsky E., Normark S., Park J.T. (1994). Bacterial cell wall recycling provides cytosolic muropeptides as effectors for beta-lactamase induction. EMBO J. 13(19):4684-4694.
9. Jacobs C., Joris B., Jamin M., Klarsov K., Van Beeumen J., Mengin-Lecreulx D., van Heijenoort J., Park J.T., Normark S., Frere J.M. (1995). AmpD, essential for both beta-lactamase regulation and cell wall recycling, is a novel cytosolic N-acetylmuramyl-*L*-alanine amidase. Mol. Microbiol. 15(3):553-559.
10. Lindberg F., Westman L., Normark S. (1985). Regulatory components in *Citrobacter freundii* ampC beta-lactamase induction. Proc. Natl Acad. Sci. USA. 82(14):4620-4624.
11. Lindquist S., Galleni M., Lindberg F., Normark S. (1989). Signalling proteins in enterobacterial AmpC beta-lactamase regulation. Mol. Microbiol. 3(8):1091-1102.
12. Lindquist S., Weston-Hafer K., Schmidt H., Pul C., Korfmann G., Erickson J., Sanders C., Martin H.H., Normark S. (1993). AmpG, a signal transducer in chromosomal beta-lactamase induction. Mol. Microbiol. 9(4):703-715.
13. Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000). GenomeExplorer: software for analysis of complete bacterial genomes. Mol. Biol. 34:222-231.
14. Normark S. (1995). β-Lactamase induction in gram-negative bacteria is intimately linked to peptidoglycan recycling. Microb. Drug Resist. 1(2):111-114.
15. Nunez C., Moreno S., Cardenas L., Soberon-Chavez G., Espin G. (2000). Inactivation of the ampDE operon increases transcription of algD and affects morphology and encystment of *Azotobacter vinelandii.* J. Bacteriol. 182(17):4829-4835.
16. Overbeek R., Larsen N., Pusch G.D., D'Souza M., Selkov E., Jr., Kyrpides N., Fonstein M., Maltsev N., Selkov E. (2000). WIT: integrated system for high-throughoutput genome sequence analysis and metabolic reconstruction. Nucl. Acids Res. 28(1):123-125.
17. Ottolenghi A.C., Ayala J.A. (1991). Induction of a class I beta-lactamase from *Citrobacter freundii* in *Escherichia coli* requires active ftsZ but not ftsA or ftsQ products. Antimicrob. Agents Chemother. 35(11):2359-2365.
18. Park J.T. (1993). Turnover and recycling of the murein sacculus in oligopeptide permease-negative strains of *Escherichia coli:* indirect evidence for an alternative permease system and for a monolayered sacculus. J. Bacteriol. 175(1):7-11.
19. Park J.T. (1995). Why does *Escherichia coli* recycle its cell wall peptides? Mol. Microbiol. 17(3):421-426.
20. Park J.T. (1996). The convergence of murein recycling research with beta-lactamase research. Microb. Drug Resist. 2(1):105-112.
21. Park J.T. (2001). Identification of a dedicated recycling pathway for anhydro-N-acetylmuramic acid and N-acetylglucosamine derived from *Escherichia coli* cell wall murein. J. Bacteriol. 183(13):3842-3847.
22. Perego M., Higgins C.F., Pearce S.R., Gallagher M.P., Hoch J.A. (1991). The oligopeptide transport system of *Bacillus subtilis* plays a role in the initiation of sporulation. Mol. Microbiol. 5(1):173-185.
23. Tolg M., Schmidt H., Schierl R., Datz M., Martin H.H. (1993). Dependence of induction of enterobacterial AmpC beta-lactamase on cell-wall peptidoglycan, as demonstrated in *Proteus mirabilis* and its wall-less protoplast L-form. J. Gen. Microbiol. 139:2715-2722.
24. Tuomanen E., Lindquist S., Sande S., Galleni M., Light K., Gage D., Normark S. (1991). Coordinate regulation of beta-lactamase induction and peptidoglycan composition by the amp operon. Science. 251(4990):201-204.
25. Wiedemann B., Dietz H., Pfeifle D. (1998). Induction of beta-lactamase in *Enterobacter cloacae.* Clin. Infect. Dis. 27: S42-S47.

BIOINFORMATICS APPROACH TO ANALYSIS OF REGULATION OF AROMATIC AMINO ACIDS BIOSYNTHESIS IN *BACILLUS/CLOSTRIDIUM* GROUP

[1] Panina E., [1,2] Vitrehschak A., [1] Mironov A., [1] Gelfand M.

[1] Branch of Corporation Integrated Genomics, Inc., postbox 348, 117333, Moscow, Russia
[2] Institute of Problems for Information Transmission, RAS
e-mail: katya@ekpanina.mccme.ru
*Corresponding author

Key words: *aromatic amino acids, comparative genomics, T-box, TRAP, transcription, regulation, ABC transporter*

Resume

Motivation: While regulation of aromatic amino acids biosynthesis (AAAB) has been intensely studied in *Bacillus subtilis*, little is known about the mechanisms of regulation in other members of the *Bacillus*/*Clostridium* group. Since most species in this group are dangerous human pathogens, e.g. *Bacillus anthracis* and *Staphylococcus aureus*, the theoretical research in this area is highly important.
Results: We have applied the comparative genomics approach to analysis of regulatory patterns involved in AAAB in the *Bacillus/Clostridium* group. We demonstrate the variability of DNA and RNA regulation of orthologous genes in different species. We describe a new type of transcriptional regulation of DAHP synthase and shikimate kinase genes in the *Streptococcus* and *Lactococcus* species, and new candidate T-boxes upstream of AAAB genes in the analyzed genomes. Finally, we identify a candidate tryptophan transporter in the *Streptococcus*, *Lactococcus*, *Enterococcus,* and *Desulfitobacterium* species.

Introduction

Biosynthesis of three aromatic amino acids starts with the common pathway leading from phosphoenolpyruvate (PEP) and erythrose 4-phosphate (E4P) through 3-deoxy-D-arabino-heptulosonate-7-phosphate (DAHP) and shikimate to the chorismic acid (genes *aroA, aroB, aroC, aroI, aroD, aroE,* and *aroF* in *Bacillus subtilis*). Then, the pathway divides into the terminal pathways, specific for each aromatic amino acid (genes *trpE, trpG, trpD, trpC, trpF, trpB,* and *trpA* for the tryptophan production; *aroA, pheA, pheB, aroH, tyrA, hisC,* and *aspB* for the phenylalanine and tyrosine production in *Bacillus subtilis*).
In gram-positive bacteria, no transcriptional regulation of AAAB has yet been experimentally discovered. However, Terai et al. (2001) have identified PCEs (phylogenetically conserved elements) upstream of *aroA* genes in *B. subtilis* and *Bacillus halodurans*, and upstream of *aroF* genes in the *B. subtilis* and *Bacillus stearothermophilus*, which might play a role in the transcriptional regulation of AAAB in *Bacillus* species. The RNA regulation of this pathway in gram-positive bacteria involves the RNA-binding protein TRAP that regulates transcription and translation of the *trpEDCFBA* operon and translation of the *trpG* and *yhaG* genes in *B. subtilis*, the latter encoding a candidate tryptophan-specific permease (Bobitzke, Gollnick, 2001). The other type of the RNA-level regulation is presented by T-boxes that regulate transcription of the *trpEGDCFBA* operon in *Lactococcus lactis* (Fig. 1).
We have previously applied the comparative genomics approach to the analysis of DNA- and RNA-level regulation of AAAB in γ-proteobacteria (Panina et al., 2000). Here, we apply the same approach to the analysis of regulatory patterns involved in this pathway in gram-positive bacteria of the *Bacillus/Clostridium* group: *Bacillus*, *Clostridium, Streptococcus, Enterococcus, Lactococcus, Staphylococcus, Lysteria,* and *Desulfitobacterium* species.

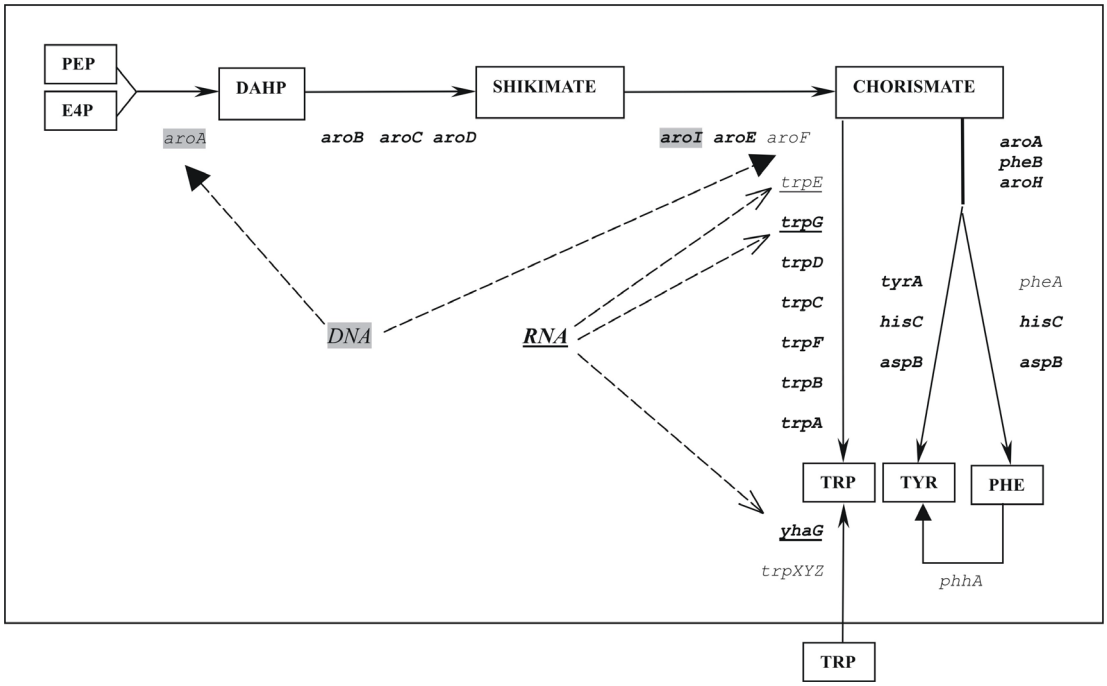


Fig. Genes encoding the enzymes of the aromatic amino acids biosynthesis pathway, their regulation, and the transporters for tryptophan (*yhaG* is a known transporter, and *trpXYZ* is predicted to be a tryptophan transporter in this study). The known regulation is shown by dotted lines: filled arrows, DNA-level regulation and PCEs; empty arrows, RNA level regulation, TRAP (underlined), and T-boxes (bold). Candidate regulation found in this study: shaded, new type of transcription regulation and bold, new T-boxes.

Materials and Methods

Complete genome sequences of *Bacillus subtilis, B. halodurans, Streptococcus pneumoniae, Lactococcus lactis, Enterococcus faecalis, Streptococcus pyogenes, Clostridium acetobutilicum, Staphylococcus aureus,* and Listeria monocytogenes were downloaded from GenBank (Benson et al., 2000). Partially sequenced genomes of *Bacillus stearothermophilus, Streptococcus mutans, Clostridium difficile,* and *Desulfitobacterium halfniense* were extracted from the ERGO Database (http://wit.mcs.anl.gov/WIT2/). Partially sequenced genome of *Enterococcus faecium* was obtained from the DOE Joint Genome Institute site (http://www.jgi.doe.gov); and partially sequenced genome of *Bacillus anthracis* was obtained from the WWW site of the Institute for Genomic Research (http://www.tigr.org).

Profiles for signal recognition were constructed as described in (Panina et al., 2001). Positional nucleotide weights in these profiles are defined as

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

where N(b,k) is the count of nucleotide b at position k. The score of the candidate site is calculated as the sum of the respective positional nucleotide weights:

$Z(b_1 \ldots b_L) = \Sigma_{k=1\ldots L} W(b_k,k)$, where k is the length of the site.

Genomic analyses (protein similarity searches using Smith-Waterman algorithm, analysis of orthology, and identification of candidate signals in the genome sequences) were done using GenomeExplorer (Mironov et al., 2000). Searches for RNA secondary structure sites were performed using RNApattern.

Results and Discussions

The pathway: genes and operons. While the backbone of AAAB pathway is conserved in most bacterial species, we have identified some steps that vary within the analyzed group. First, the complete genomes of *S. pyogenes* and *E. faecalis* lack genes for terminal tryptophan pathway. Second, in *S. pyogenes* there are no homologs of *pheA* and *tyrA* genes from the terminal phenylalanine and tyrosine pathways, respectively. Third, in *S. pneumoniae, S. mutans,* and *L. lactis,* there are no homologs of the DAHP synthase gene *aroA* of *B. subtilis*, while there are two genes homologous to DAHP synthases from gram-negative bacteria. Next, in *B. anthracis* and *D. halfniense*, there is a homolog of the *phhA* gene previously identified only in a number of Proteobacteria and eukaryotes. PhhA catalyzes the conversion of phenylalanine to tyrosine. Finally, the *Bacillus, Streptococcus,* and *Clostridium* genomes, excluding only *B. anthracis*, have one copy of *trpG* gene that functions both in the tryptophan and folate biosynthesis, whereas *L. lactis, S. aureus, B. anthracis,* and *D. halfniense* have two paralogous copies of this gene.

The operon structure of AAAB genes varies significantly. The only conserved feature is the *trpE*(*G*)*DCFBA* operon, which is either absent or present as a whole. The only exception is the *trpG* gene that lies either in the *trp* operon (in

S. pneumoniae, S. mutans, and *C. acetobutilicum*), or in the folate biosynthesis operon (in *B. subtilis, B. halodurans, B. stearothermophilus, C. difficile,* and *S. pyogenes*). In *L. lactis, S. aureus, B. anthracis,* and *D. halfniense*, where there are two copies of *trpG* gene, one copy lies in the *trp* operon, whereas the other one is co-localized with the folate biosynthesis genes. Thus, we propose that the duplicated enzymes have acquired narrow specificity for tryptophan (RLLX01504, RSA03401, RZC03347, RDHA05110) and folate (RLLX01346, RSA02493, RZC04175, RDHA04984) production, respectively.

DNA-level regulation. Pairs of DAHP-synthase genes of *S. pneumoniae*, *S. mutans,* and *L. lactis,* encoding homologs to gram-negative, rather than gram-positive enzymes, form operons in *S. pneumoniae*, *S. mutans*, but are located separately in *L. lactis.* We have found a conserved 14-bp sequence ATGGAGGCANATAA upstream of the DAHP synthase operons in *S. pneumoniae* and *S. mutans*, and upstream of both DAHP synthases genes in *L. lactis.* Moreover, a similar sequence was found in the upstream regions of the shikimate kinase genes in all the three species. Notably, the reactions catalyzed by shikimate kinase and DAHP synthase are the only two irreversible steps within the common pathway of AAAB, and only these genes of the common pathway are regulated at the transcriptional level in γ-proteobacteria. Thus, we propose that the new conserved sequence plays a role in transcriptional regulation of DAHP synthase and shikimate kinase genes in *Streptococcus* and *L. lactis* genomes.

We have constructed the profile based on the PCEs described in (Terai et al., 2001). Using this profile we have found one more candidate site ACTTAAccaCGTT upstream of the *aroF* gene in *B. halodurans.*

RNA-level regulation. A number of T-boxes were found upstream of genes involved in AAAB. In particular, tyrosine-specific T-boxes were found upstream of the *aroA*, *aroF,* and *phhA* genes in *B. anthracis*; tryptophan-specific T-boxes were found upstream of the *trp* operons in *B. anthracis*, *S. pneumoniae, S. mutans, L. lactis, C. acetobutilicum, S. aureus,* and *L. monocytogenes.* We have also found a phenylalanine-specific T-box upstream of the *pheA* gene in *D. halfniense.*

Candidate TRAP-binding sites were found upstream of the *trp* operons and *trpG* genes in *B. halodurans* and *B. stearothermophilus*.

Interchange of regulatory systems. So far, there seem to be four types of regulation of AAAB in *Bacillus/Clostridium* group. The most general is the T-box-dependent transcriptional regulation, which is present in all the studied species. Another type of the RNA-dependent transcriptional regulation, TRAP-mediated regulation, is unique to the *Bacillus* group except for *B. anthracis*, which lacks the TRAP protein. In *B. subtilis*, *B. halodurans,* and *B. stearothermophilus*, TRAP regulates transcription of the *trp* operon, which is regulated by tryptophan-specific T-boxes in all the other species. The third type of regulation, PCEs, is also specific of *B. subtilis*, *B. halodurans* and *B. stearothermophilus*, where it appears to regulate the transcription of DAHP synthase and chorismate synthase genes. In *B. anthracis,* the same genes are regulated by tyrosine-specific T-boxes, while in *S. pneumoniae*, *S. mutans,* and *L. lactis,* DAHP synthases as well as shikimate kinases genes are under the control of the fourth type of transcriptional regulation identified in this study.

Transporters of aromatic amino acids. The only known tryptophan transporter in the *Bacillus/Clostridium* group is YhaG of *B. subtilis*, whose translation is regulated by TRAP protein. We have found orthologs of the *yhaG* gene in *B. stearothermophilus*, *C. acetobutilicum,* and *C. difficile*; however, no homologs of *yhaG* could be observed in the genomes of *E. faecalis* and *S. pyogenes,* which lack the tryptophan biosynthesis pathway, and thus, should transport tryptophan from the environment. We have identified tryptophan-specific T-boxes upstream of the *yhaG* orthologs in both *Clostridium* species; in *B. stearothermophilus,* the upstream region of this gene is not sequenced yet.

We have found a new candidate tryptophan ABC transporter, named *trpXYZ*, in the genomes of *S. pneumoniae*, *S. mutans*, *S. pyogenes, S. equi*, *E. faecalis*, *E. faecium, B. stearothermophilus, D. halfniense, B. cepacia,* and *M. loti* (the last two are α-proteobacteria). The genes in *S. pneumoniae genome* are *SP1069, SP1070, and SP1071*. *trpXYZ* is presented in three copies in the genome of *D. halfniense*, and two of them have tryptophan-specific T-boxes in the upstream regions. Besides, *trpXYZ* is preceded by a tryptophan-specific T-box in *S. pneumoniae.* Moreover, *trpXYZ* is co-localized with the *aroD* gene in *E. faecium,* and with gene encoding enzymes of the tryptophan degradation kynurenine pathway in *M. loti.* These observations allow us to ascribe the tryptophan specificity to this transporter.

Acknowledgements

We are grateful to Dmitry Rodionov for useful discussions. This study was partially supported by grants from INTAS (99-1476) and HHMI (55000309).

References

1. Bobitzke P., Gollnick P. (2001). Posttranscriptional initiation control of tryptophan metabolism in *Bacillus subtilis* by the trp *RNA*-binding attenuation protein (TRAP), anti-TRAP, and RNA structure. J. Bacteriol. 183, 5795-5802.
2. Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Rapp B.A., Wheeler D.L. (2000). GenBank. Nucl. Acids Res. 28, 15-18.
3. Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000). GenomeExplorer: software for analysis of complete bacterial genomes. Mol. Biol. 34, 222-231.
4. Panina E.M., Vitreschak A.G., Mironov A.A, Gelfand M.S. (2001). Regulation of aromatic amino acid biosynthesis in gamma-proteobateria. J. Mol. Microbiol. Biotechnol. 3, 529-543.
5. Terai G., Takagi T., Nakai K. (2001). Prediction of co-regulated genes in *Bacillus subtilis* on the basis of upstream elements conserved across three closely related species. Genome Biol. 2, research 0048.1–0048.12.

REGULATION OF THE HEAT SHOCK RESPONSE OF β-, γ- AND ε-PROTEOBACTERIA

[1] Permina E.A., [1, 2] Gelfand M.S.*

[1] GosNII Genetika, Moscow, Russia, e-mail: epermina@mail.ru
[2] Integrated Genomics - P. O. Box 348, 117333, Moscow, Russia

Key words: *heat shock, regulation, thioredoxin, HrcA/CIRCE, sigma factors*

Resume

Motivation: Our motivation was to study the regulation of heat shock response in the β-, γ- and ε-proteobacteria by the means of comparative analysis.
Results: We have predicted a number of new heat shock regulon members and suggested the function for them. Also we suggest the cross-regulation between the two major heat shock regulators – HrcA and σ^{32}.

Introduction

During heat shock the main strategy of an organism is defence from denatured proteins. This is done by chaperones that refold and proteases that cut abnormal proteins.
The heat shock response in bacteria often has complex regulation that depends on more than one regulator. In *E. coli* there are three sigma factors implicated in the heat shock response. Some other proteobacteria have in addition the HrcA/CIRCE system.
σ^{32} is widely distributed among γ-proteobacteria and the promoter sequences it recognizes are strongly conserved. Its consensus CTTGAAA-N16-CCCCAT is still recognizable in β-proteobacteria and has undergone some changes in α-proteobacteria. The σ^{32} regulon is rather large and includes genes encoding chaperones, proteases and proteins with predicted mixed activity.
The HrcA/CIRCE system is distributed much wider than σ^{32}. CIRCE is a palindrome with highly conserved wings TTAGCACTC-N9-GAGTGCTAA. It is one of the most conserved signal sequences in the eubacterial world and is found in various representatives of Cyanobacteria, Spirochaetes, Firmicutes, Proteobacteria and Chlamidiae. The repressor protein HrcA that binds to this palindrome shares its unusual conservancy and is clearly recognizable from *Thermotoga maritima* through *Bacillus subtilis* and γ-proteobacteria. Unlike σ^{32}, in most genomes HrcA regulates only chaperones. Regulation of *groESL* by HrcA seems to be obligate if an organism does have the HrcA/CIRCE system.

Data and Methods

The comparative approach to the analysis of transcriptional regulation in bacterial genomes is based on the assumption that sets of genes regulated by orthologous transcription factors are conserved in related genomes. This technique can be used when genomes of related organisms are avaliable although the relation should not be too close (f.e. γ-proteobacteria). Thus the candidate sites occurring upstream of orthologous genes are true, whereas false positives are scattered at random.
The positional nucleotide weights in these profiles were defined as $W(b,k) = log\,[N(b,k) + 0.5\,] - 0.25\,\Sigma_{i=A,C,G,T} \log\,[\,N(i,k) + 0.5\,]$, where $N(b,k)$ denoted the count of nucleotide b at position k. The score of a L-tuple candidate site was calculated as the sum of the respective positional nucleotide weights $Z(b_1 \ldots b_L) = \sum_{k=1 \ldots L} W(b_k, k)$. The base of the logarithm was chosen such that the distribution of the Z-score on random L-mers is Gaussian. Thus the Z-score can be used to assess the significance of an individual site (Gelfand, 2000).
The cutoff score for potential CIRCE sequences was 12.71 (5 substitutions). We took σ^{32} promoters that have score over 6.00 (less than 5% of the genome).
The following genomes were considered: *Campylobacter jejuni*, *Escherichia coli, Haemophilus influenzae* Rd, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Neisseria meningitidis* MC58, *Pseudomonas aeruginosa* PAO1, *Vibrio cholerae*, *Yersinia pestis*, *Xylella fastidiosa* extracted from GenBank, as well as partially sequenced genomes of *Bordetella pertussis, Salmonella typhi,* and *Nitrosomonas europaea* obtained from the TIGR Web site ({{http://www.tigr.org}}) and *Bordetella bronchiceptica, Bordetella parapertussis, Burkholderia cepacia, Burkholderia pseudomallei, Methylobacillus flagellatus,* and *Ralstonia eutropha* obtained from ERGO ({{http://wit.integratedgenomics.com/igwit}}). The profiles for CIRCE sites and σ^{32} promoters were constructed using the samples from (Narberhaus, 1999; Gross, 1996), respectively.
Genomic analyses (genomic scale similarity searches, site searches using profiles etc.) were done using GenomeExplorer (Mironov et al., 2000) ClustalX 1.5 (Thompson et al., 1997). Signal profiles were constructed using SignalX (Mironov et al., 2000). Database similarity searches were done by BLAST (Altschul et al., 1997) at the NCBI Web site ({{http://www.ncbi.nlm.nih.gov/BLAST}})

Results

We have demonstrated that in some β-proteobacteria the *rpoH* genes encoding σ^{32} have strong CIRCE sites in upstream regions (Table 1). Several genomes, in particular, *Methylobacillus flagellatus KT* and *Bordetella parapertussis* may provide an example of cross-regulation of different regulators responding to the same environmental stimulus. Indeed, the upstream regions of the *hrcA* genes in these genomes have strong candidate σ^{32} promoter sequences (Table 2).

Table 1. CIRCE sites upstream of *rpoH* and *groESL* genes in β-proteobacteria.

Genome	ERGO ID	Gene	Predicted HrcA binding site		
			pos	score	sequence
R. eutropha	RREU04047	*rpoH*	-97	18.00	TTAGCACTC-(9)-GAGTGCTAA
B. pertussis	RBP01279	*rpoH*	-73	18.00	TTAGCACTC-(9)-GAGTGCTAA
B. parapertussis	RBPA04410	*rpoH*	-92	18.00	TTAGCACTC-(9)-GAGTGCTAA
B. pseudomallei	RBPS04433	*rpoH*	-87	18.00	TTAGCACTC-(9)-GAGTGCTAA
N. europaea	RNE01139	*rpoH*	-84	16.67	TTAGCACTC-(9)-GAGTGCTAg
M. flagellatus	RMFL02417	*rpoH*	-56	14.01	cTAGCACaC-(9)-GAGTGCTAg
Methylovorus sp. SS1	AF177466	*rpoH*	?	18.00	TTAGCACTC-(9)-GAGTGCTAG

Table 2. Predicted σ^{32}-dependent promoters upstream of heat shock genes in β-proteobacteria.

Genome	ERGO ID	Gene	σ^{32}-dependent promoter		
			pos	score	sequence
B. pertussis	RBP01020	hrcA	-27	6.23	gTTGAAA-(15)-gCtCAT
B. parapertussis	RBPA01257	hrcA	-25	6.23	gTTGAAA-(15)-gCtCAT
M. flagellatus	RMFL01214	hrcA	-39	6.22	aTTGAAt-(14)-CCtCAT

Positional analysis revealed a potential new member of the heat shock response regulon. In the genomes of *Bordetella* species and *R. eutropha,* the *grpE* gene regulated by σ^{32} is followed by a gene encoding a hypotetical thioredoxin. Moreover, many γ-proteobacteria, in particular *E. coli*, *K. pneumoniae, S. typhi, Y. pestis, V. cholerae* and *X. fastidiosa* also have thioredoxin-resembling genes with candidate σ^{32} promoters (Table 3). The latter genes are not homologous to the predicted thioredoxin genes from β-proteobacteria. In *X. fastidiosa,* the thioredoxin-like protein has double potential regulation by HrcA/CIRCE and σ^{32}.

Table 3. Predicted members of the σ^{32} regulon in γ-proteobacteria

Genome	gene name	position	score	site sequence
Escherichia coli	b0492	-23	7.41	gTTGAAg-(13)-CCCCAT
Salmonella typhi	RTY00345	-57	7.41	gTTGAAg-(13)-CCCCAT
Yersinia pestis	RYP00871	-58	7.41	gTTGAAg-(13)-CCCCAT
Vibrio cholerae	VC0977	-41	6.94	CTTGAgt-(14)-CCCCAT
Xylella fastidiosa	XF2174	-108	6.15	CTTGtAt-(13)-CaCCAT

The ε-subdivision stands apart from other proteobacteria. The *hrcA*-like genes of *H. pylori* and *C. jejuni* are closer to the genes annotated as *hrcA* in *Chlamidiae* but still are only distantly related to other members of the family. The similarity of *C. jejuni* and *B. subtilis hrcA* genes cannot be seen by BLAST alone: the E-value of comparison of *hrcA* from *H. pylori* and *C. jejuni* is ~ 10^{-16}, whereas the comparison of *hrcA* of *C. jejuni* with *hrcA* of *B. halodurans* yields E-value of 0.054. However, since this gene is likely to form one operon with *grpEdnaKJ,* we can assume that it plays the same role as standard *hrcA*. We have identified a common palindrome of 27 nucleotides upstream of *H. pylori* and *C. jejuni hrcA*-like genes and the *groESL* operons (Table 4). This palindrome (the consensus is AAAATTTAGTC aaata GACTAAATTTT, upper-case for palindromic bases, lower-case for nonpalindromic gap) can be the binding signal of the unusual HrcA of ε-proteobacteria.

Table 4. Candidate HrcA sites in *H. pylori* (the training set) and *C. jejuni*.

Genome	Gene name	Position	Score	Sequence
H. pylori	HP0011 (groES)	-155	6.16	AAAcTTgAtaCAAATAGACTtAATaaT*
H. pylori	HP0111 (hrcA)	-118	6.31	tAgATTTAGTgAtATAGACTAAAcTTT*
C. jejuni	hrcA	-129	6.40	AAAcTTTAGTCAtATTGACTAAATaaa
C. jejuni	groES	-135	6.15	tAAcTTTAGTCTATaAaACTAAAcTTT

* The sequence was used in the training set.

Discussion

In γ-proteobacteria we have observed conserved σ^{32}-dependent regulation of b0492 homologs which are similar to thioredoxines (f. e. *trxA* of *M. leprae* - $6e^{-13}$) and disulphide isomerases (NC_003450 – $2e^{-13}$). The conservation of the

potential signal upstream of these genes indicate that they may play some particular role in heat shock response. The positional analysis revealed co-location of the heat shock gene *grpE* with another putative thioredoxin gene in some β-proteobacteria (data not shown). One possible role of these proteins from β-proteobacteria during the heat shock could be maintaining the correct folding of proteins by acting on cysteine bridges as they have intact active site.
Another important result is prediction of cross-regulation of σ^{32} and HrcA in *Bordetella* and *Methylobacillus* and regulation of the *rpoH* gene by HrcA which used to be thought to regulate only chaperone genes.

Acknowledgements

This study was partially supported by grants from INTAS (99-1476), HHMI (55000309), and RFBR (0015-99363). We are grateful to Andrei Mironov, Alexandra Rakhmaninova, Eugene Koonin and Ekaterina Panina for useful discussions.

Reference

1. Altschul S., Gish W., Miller W., Myers E., Lipman D. Basic local alignment search tool. J. Mol. Biol. 1990. 215:403-410.
2. Gelfand M.S. Novichkov P.S., Novichkova E.S., Mironov A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. Brief Bioinform. 1, 357-71.
3. Gross C.A. Function and Regulation of the Heat Shock Proteins. Escherishia coli and Salmonella, Cellular and Molecular Biology ASM. Neidhardt F.C. Washington, DC. Press, 1996. V. 1. P. 1400-1412.
4. Mironov A.A., Vinokurova N.P., Gelfand M.S. (2000) Software for analysis of bacterial genomes. Mol. Biol. 34, 222-231.
5. Narberhaus F. (1999) Negative regulation of bacterial heat shock genes. Mol. Microbiol. 31, 1-8.
6. Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F., Higgins D.G. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl. Acids Res. 25, 4876-82.

PURINE REGULON OF GAMMA-PROTEOBACTERIA

Ravcheyev D.A., Gelfand M.S. *, Mironov A.A., Rakhmaninova A.B.

State Scientific Center of Biotechnology "NII Genetika", Moscow, 113545, Russia
*e-mail: misha@imb.imb.ac.ru

Key words: *comparative genomics, PurR, purine regulon, gamma-proteobacteria*

Resume

Motivation: Purine regulon is one of best studied regulons in *E. coli*. However, purine regulation in the other bacteria is poorly characterized.
Results: Six genomes of gamma-proteobacteria, *Escherichia coli, Salmonella typhi, Yersinia pestis, Haemophilus influenzae, Pasteurella multocida,* and *Vibrio cholerae* were studied. Candidate binding sites of the purine repressor are conserved upstream of genes for IMP synthesis from ribose 5-phosphate, genes implicated in one-carbon-compound metabolism and genes for the set of transport proteins. Furthermore, conserved PurR sites were detected upstream of genes involved in the synthesis of pyrimidines and ribose methabolism.

Introduction

The purine regulon of *E. coli* includes genes whose transcription is regulated by the purine repressor PurR. PurR is a dimer of 38-kDa subunits. The N-terminal DNA binding domain contains a helix-turn-helix motif which contacts the maijor groove. PurR binds to a 16-bp palindrome in the control region thus preventing initiation or elongation of transcription. The consensus sequence for the PurR binding site was determined as ACGCAAACGTTTGCGT.
The purine regulon contains genes for the *de novo* purine biosynthesis (*prsA, purF, purHD, purT, purMN, purL, purEK, purC, purB, purA* and *gua BA*), genes involved in the pyrimidine synthesis (*pyrC, pyrD, codBA*), genes implicated in the one-carbon-compound metabolism (*glyA, gcvTHP*), and some nitrogen metabolism genes (*speAB* and *glnB*). The purine regulation of all of these genes had been demonstrated in experiment (Zalkin, Nygaard, 1996).

Methods

Complete genome sequences of *Escherichia coli*, *Yersinia pestis*, *Haemophilus influenzae, Pasteurella multocida* and *Vibrio cholerae* were extracted from GeneBank. The partial sequence of *Salmonella typhi* was extracted from the Sanger Institute web site (http://www.sanger.ac.uk/).
For inverstigation of the regulon composition, the comparative approach was used. This approach is based on the assumption that sets of genes regulated by orthologous transcription factors are conserved in related genomes. This technique can be used when genomes of related organisms are avaliable, although the relation should not be too close. Candidate sites were predicted using positional nucleotide weight matrices (Mironov et al., 1999). A site was accepted if it was observed in the (-300 … +100) region of orthologous genes in more that one genome. Genes were considered to belong to one operon if they were transcribed in the same direction and the intergenic distance did not exceed 100 nucleotides.

Results and Discussion

Purine sites upstream of the *purR* gene were found only in Enterobacteriacea*e*.
All genomes retain the regulation of the genes responsible for the IMP synthesis: *purL, purEK, cvpApurF, purMN, purHD, purB*.
There are no purine sites upstream of the *purT* gene in *H. influenzae* and *P. multocida*. This gene is directly involved in the purine biosynthesis, but its function is redundant with that of *purN.*
purC is subject to non-orthologous gene displacement by distant homologs in Pasteurellaceae and *V.cholerae*. It is preceeded by PurR sites in all genomes except *Y. pestis.*
Genes involved in AMP and GMP from IMP synthesis, *purA* and *guaAB* retained purine regulation in a fraction of *E. coli* and *S. typhi*.
The operon *upp-uraA*, encoding uracil transporter and uracil phosphoribosyltransferase, retains its structure and PurR sites in all genomes except *V. cholerae.* In *V. cholerae* this operon is disrupted but each gene has its own PurR site.
Purine sites upstream of the *gcvTHP* operon were found only in Enterobacteriaceae. However in other genomes were are PurR sites upstream of *folD,* which is a functional analog of *gcvT*. Thus, genes reponsible to the folate-associated one-carbon-compound methabolism are obligatory members of the purine regulon.
glyA, encoding serine hydroxymethyltransferase retains PurR sites in all Enterobacteriaceae and in *V.cholerae* genomes. In *H. influenzae* and *P. multocida* this gene lies downstream of the *purD* gene and presumably belongs to the PurR-regulated operon *purHDglyA*.

The gene for 3-phosphoglycerate dehydrogenase, *serA*, has a PurR site in Enterobacteriaceae and *V.cholerae* genomes. In Pasteurellaceae, *serA* is preceded by *rpiA* encoding ribose 5-phosphate isomerase A. These genes presumably form an operon and in *H. influenzae* and *P. multocida* a PurR site was observed upstream of *rpiA* gene.
Thus the purine regulon should be supplemented by genes *folD* in Pasteurellaceae and Vibrionaceae, *rpiA* in Pasteurellaceae, and *upp, uraA* and *serA* in all considered genomes.
Our resultes demonstrate that changes in the purine regulon composition are specific for different groups of gamma-proteobacteria and result from divergence in evolution of each taxonomy group.
The comparative approach used in this study allows one to determine the core of a regulon while species-specific regulated genes can be last. However, increasing the number of studied genome results in finding this "rare" members of regulons.
To inverstigate the differences in structure of the PurR binding signal, positional nucleotide weight matrices and sequence logos for *E. coli*, *Y. pestis*, *H. influenzae*, *P. multocida* and *V. cholerae* were constructed. At that, all predicted PurR sites in each genome were used as learning samples.
There was no significant differences in the structure of *E. coli* and *Y. pestis* PurR binding sidnals. In *H. influenzae*, *P. multocida* and *V. cholerae*, substitutions at symmetrical positions 2 and 15 were observed.

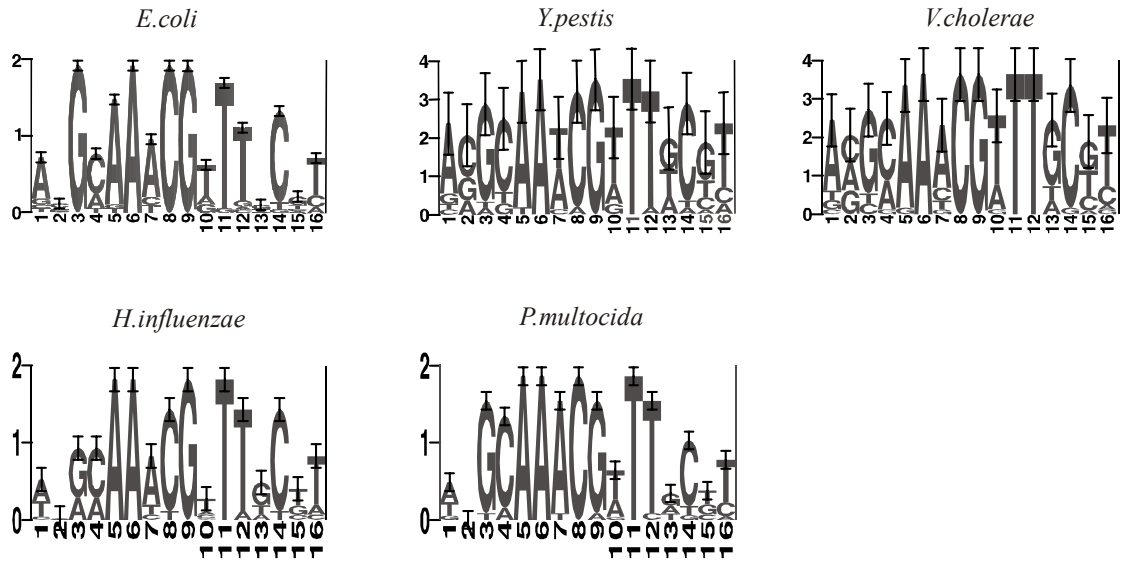


Fig. Sequence logos for PurR sites from *E. coli, Y. pestis, H. influenzae*, *P. multocida* and *V. cholerae*.

Acknowlegements

We are grateful to E.Panina, E.Permina, D.Rodionov and O.Laikova for useful discussion.
This work was partially supported by grants from INTAS (99-1476), the Howard Hughes Medical Institute (55000309), and the Russian Fund of Basic Research (00-15-99362).

References

1. Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. Nucl. Acids Res. 15. 2981-2989.
2. Zalkin H., Nygaard P. (1996) Biosynthesis of purine nucleotides. In Neighart F.C. (eds.). Escherichia and Salmonella Cellular and Molecular Biology. ASM Press, Washington, 561-579.

COMPUTATIONAL ANALYSIS OF THE BIOTIN REGULON IN BACTERIAL GENOMES

Rodionov D.A. *[1], Mironov A.A. [2], Gelfand M.S. [1,2]

[1] State Scientific Center GosNIIGenetika, Moscow, 113545, Russia, e-mail: rodionov@genetika.ru
[2] Integrated Genomics – Moscow, P.O.Pox 348, Moscow, 117333, Russia
*Corresponding author

Key words: *genome analysis, biotin biosynthesis and transport, BirA, BioY, bacteria*

Resume

Motivation: The strict control of biotin biosynthesis in *Escherichia coli* is mediated by the bifunctional BirA protein which acts both as a biotin-protein ligase and a transcriptional repressor of the biotin operon. Little is known about regulation of biotin biosynthesis in other bacteria. Thus, the complete description of the biotin regulon in bacteria as well as new functional predictions, mainly of unknown biotin transporters, are necessary.
Results: Using comparative genomics and phylogenetic analysis, we describe the biotin biosynthetic pathway and the BirA regulon in most available bacterial genomes. Existence of N-terminal DNA-binding domain in BirA strictly correlates with presence of putative BirA-binding sites upstream of biotin operons. The predicted BirA-binding sites are well-conserved among various eubacterial and archaeal genomes. The possible role of hypothetical genes *bioY* and *yhfS-yhfT*, newly identified members of the BirA regulon, in biotin metabolism is discussed. Based on analysis of co-occurrence of the biotin biosynthetic genes and *bioY* in complete genomes, we predict involvement of the transmembrane protein BioY in biotin transport. Different non-orthologous substitutes of the *bioC*-coupled gene *bioH* from *E. coli*, observed in several genomes, possibly represent existence of different pathways for the pimeloyl-CoA biosynthesis. Another interesting result of analysis of operon structures and BirA sites is that some biotin-dependent carboxylases from *Rhodobacter capsulatus,* actinomycetes and archaea are possibly co-regulated with BirA. BirA is the first example of a transcriptional regulator with conserved binding signal in eubacteria and archaea.

Introduction

Biotin (vitamin H) is an essential cofactor for a class of important metabolic enzymes, biotin carboxylases and decarboxylases. Biotin biosynthetic pathway is widespread among microorganisms (Fig. 1). Genes encoding biotin transporters have not been identified in bacteria until now. The operon organization of the biotin biosynthetic genes differs between *E. coli* and bacilli. *E. coli* has the *bioBFCD* operon located divergently with the *bioA* gene and the single *bioH* gene. In contrast, *B. subtilis* has only the *bioWAFDBI* operon.
The biotin operon of *E. coli* is negatively regulated by biotin and the bifunctional protein BirA. The biotin-protein ligase BirA mediates biotinylation of acetyl-CoA carboxylase via a two-step reaction. Firstly, the adenylate of biotin is synthesized from substrates biotin and ATP and, at the second step, transferred to a unique lysine residue on carboxylase. In addition, BirA can act as a repressor of transcription when it has the N-terminal DNA-binding domain (D-b-BirA).When biotin is unclaimed, two generated BirA-biotinyl-5`-AMP monomers bind cooperatively to the *bioO* operator between the divergent *bioA* and *bioBCDF* operons and repress transcription in both directions. The BirA protein is composed of the N-terminal DNA-binding (D-b) domain containing a helix-turn-helix (HTH) structure, the central domain, and the C-terminal domain. The BirA protein of *B. subtilis* has a similar structure and also can act as the repressor of the *bioWAFDBI* operon (Bower et al., 1996). Recently, two new BirA-regulated operons of unknown function, *yhfUST* and *yuiG*, were detected in *B. subtilis* by expression microarray analysis. Imperfect palindromic sequences, which are partially similar to the *bioO* operator from *E. coli,* were found upstream of the BirA-regulated operons from *B. subtilis*, *B. sphaericus* and *Kurthia* sp.

Implementation and Results

Using the global analysis of the BirA proteins and DNA-binding sites in available bacterial genomes, we have found that the BirA regulon is widely distributed in eubacteria and archaea. A correlation exists between the presence of D-b-BirA and finding of the BirA sites in bacterial genomes.
Based on the phylogenetic analysis of the D-b domains, we divided all D-b-BirA into two major groups, proteobacterial and non-proteobacterial (Fig. 2). Consistent with this, we constructed two different recognition rules (profiles) for the BirA sites using the sets of upstream regions of the biotin biosynthetic genes from various genomes. The BirA profile for proteobacteria (with consensus 5`-tTGTaAACC-N14..16-GGTTtACAa-3`, where strongly conserved positions are shown in capitals) is more strict than that for non-proteobacteria (5`-wwTGTtAAC-N14..16-GTTaACAww-3`, where ‘w’ stands for A or T).

Then we used the constructed profiles to detect new candidate members of the BirA regulons in the genomes containing D-b-BirA. Proteobacteria possess only one strong BirA site per genome occurring upstream of the biotin biosynthetic operon. However, most Gram-positive bacteria and some archaea have multiple BirA sites located upstream of biotin biosynthetic genes, *bioY* and other new genes of the BirA regulon (Table). Here we predict that *bioY* encodes a biotin transporter since this is a sole BirA-regulated gene in bacteria without biotin biosynthetic genes. Two other BirA-regulated genes, *yhfS* and *yhfT*, were also found in several bacterial genomes.

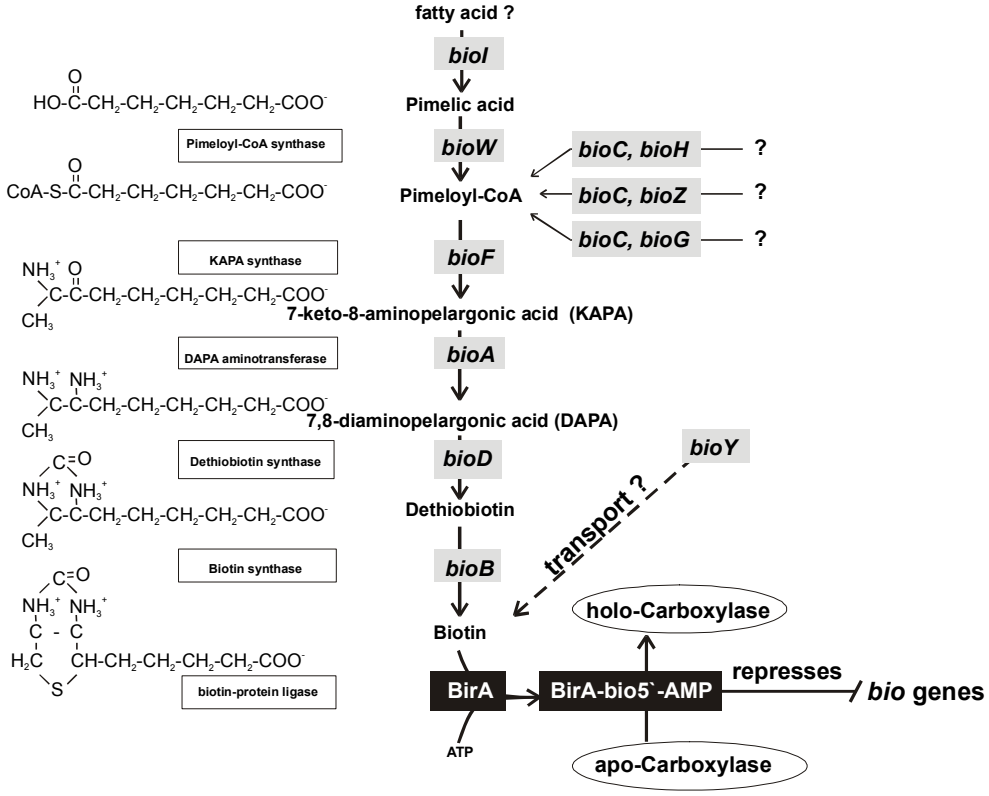


Fig. 1. The biotin biosynthesis pathway in bacteria.

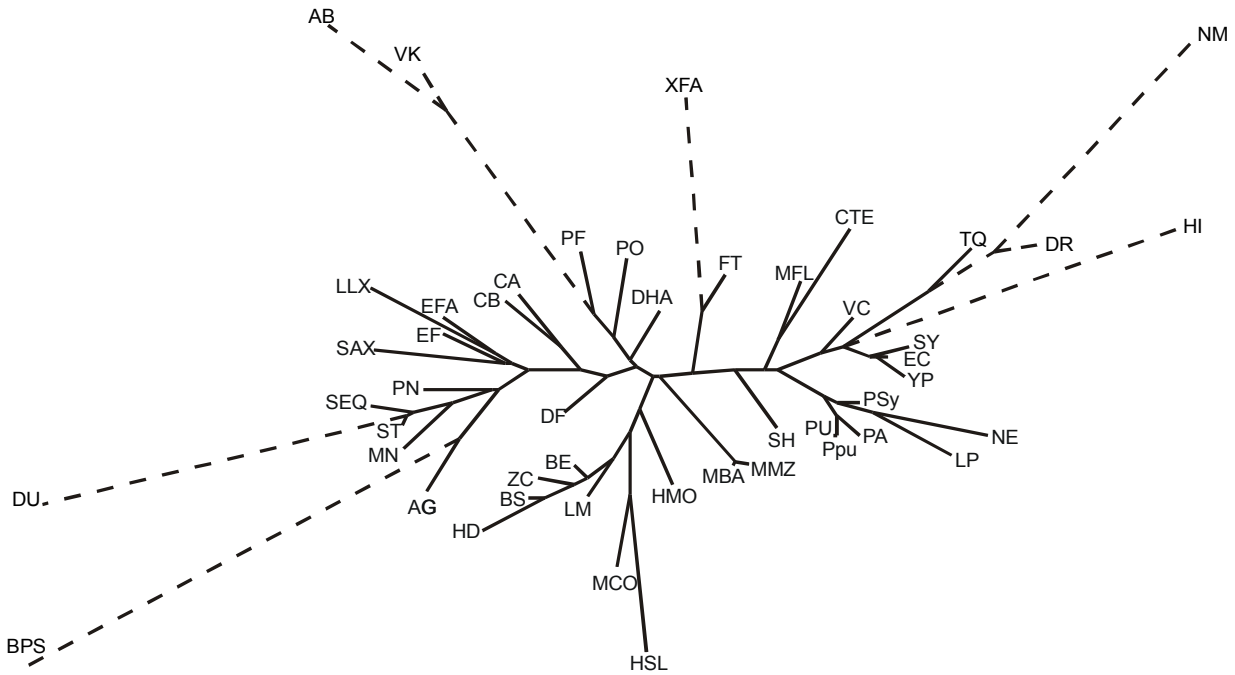


Fig. 2. The maximum likelihood phylogenetic tree for the BirA N-terminal domains. Domains which are similar to the regulatory domain of BirA from *E. coli* (containing the HTH motif) are shown in black lines. Other N-terminal domains of BirA (without HTH) are shown as an outgroup in dashed lines. The genome abbreviations are listed in Table 1.

Table. Operon structure for the biotin biosynthetic genes in prokaryotes. The genome abbreviations are given in column 'AB'. Unfinished genomes are marked by '#'. The names of taxonomic groups are given in bold. The signs '+' / '0' in the columns 'BirA D-b' and 'BirA BPL' denote existence/absence of the N-terminal regulatory domain (D-b) and C-terminal catalytic domain (BPL) of BirA, respectively; '-' denotes N-terminal BirA domain not similar to the known regulatory BirA domain. Genes forming one candidate operon (with spacer less than 100 bp) are separated by dashes. Different loci are separated by slashes. Direction of transcription in divergons is shown by angle brackets. Predicted BirA sites are denoted by '$'. The contig ends are shown by square brackets. Bio(GC) is the fusion of the *bioG* and *bioC* genes. The other genes of unknown function are denoted by *X*.

Genome	*AB*	BirA		Biotin biosynthetic genes	Biotin transporters
		D-b	BPL		
1	2	3	4	5	6
Proteobacteria					
Caulobacter crescentus	**CO**	0	+	*bioB / bioA <> bioF-bioD / bioC*	
Sinorhizobium meliloti	**SM**	0	+	*bioC*	*cbiO-cbiQ-bioY-yhfT-yhfS*
Mesorhizobium loti	**MLO**	0	+	*bioB-bioF-bioD-bioA-bioZ / bioC*	*bioY1 / bioY2-X*
Agrobacterium tumefaciens	**AT**	0	+	*bioB-bioF-bioD-bioA-bioZ / bioC*	*cbiO-cbiQ-bioY*
Brucella melitensis	**BME**	0	+	*bioB-bioF-bioD-bioA-bioZ / bioC*	*bioY1 / bioY2-X*
Rickettsia prowazekii	**RP**	0	+	none	*bioY*
Bordetella pertussis #	**BP**	0	+	*bioA <> bioF / bioB*	*cbiO-cbiQ-bioY*
Burkholderia pseudomallei #	**BPS**	-	+	*bioA-bioF-bioD-bioB / bioC*	
Nitrosomonas europaea	**NE**	+	+	*bioB-bioF-bioH-bioC-bioD / $ bioA*	
Neisseria meningitidis	NM	-	+	*bioB / bioH-bioC2 / bioF-bioG-bioC1 / bioA-bioD*	
Methylobacillus flagellatus #	**MFL**	+	+	*$ bioB-bioF-bioH-bioC-bioD / bioA-X*	
Ralstonia solanacearum	**RSO**	0	+	*bioA-bioF-bioD / X-X-bioB / bioC*	
Escherichia coli, Salmonella typhi, Klebsiella pneumoniae #, Yersinia pestis, Vibrio cholerae	**EC, TY, KP, YP, VC**	+	+	*bioA <$> bioB-bioF-bioC-bioD / bioH*	
Francisella tularensis #	**FT**	+	+	*bioA <$> bioB-bioF-bioC-bioD*	
Legionella pneumophila #	**LP**	+	+	*[bioA / [bioB-bioF-bioH-bioD / bioC*	
Haemophilus influenzae, H. ducreyi #, Pasteurella multocida, A.actinomycetemcomitans #	**HI, DU, VK, AB**	-	+	*bioA-bioF-bioG-bioC-bioD / bioB*	
Pseudomonas aeruginosa, P. putida, P. fluorescencs	**PA,PU Ppu,**	+	+	*$ bioB-bioF-bioH-bioC-bioD / bioA*	
Shewanella putrefaciens #	SH	+	+	*bioA <$> bioB-bioF-bioC-bioD / bioH*	
Thermochromatium tepidum #	**CTE**	+	+	*$ bioB-bioF-bioH-bioC] / [X-bioA*	
Xylella fastidiosa	**XFA**	-	+	*bioB / bioF-bioH / bioD / bioC / bioA*	
Helicobacter pylori	**HP**	0	+	*bioA / bioD / X-bioF / bioC / bioB-X*	
Campylobacter jejuni	**CJ**	0	+	*bioA <> bioF-bioG-bioC / X-bioD / X-bioB-X*	
Magnetococcus #	**MCO**	+	+	*$ bioF-bioH-bioC1-bioB-X-bioD / bioA / bioC2*	
Bacillus/Clostridium group					
Bacillus subtilis	**BS**	+	+	*$ bioW-bioA-bioF-bioD-bioB-bioI*	*$ bioY1 / $ bioY2-yhfT-yhfS*
Bacillus halodurans	**HD**	+	+	*$ bioB / $ bioD-bioA / $ bioF-bioH-bioC*	*$ bioY*
Bacillus stearothermophilus #	**BE**	+	+	*$ bioY1-bioD-bioA / [bioB / $ bioF*	*$ bioY2*
Bacillus cereus	**ZC**	+	+	*$ bioA-bioD-bioF-bioH-bioC-bioB*	*$ bioY1 / $ bioY2-yhfT-yhfS*
Clostridium acetobutylicum	**CA**	+	+	*$ bioY1-bioD-bioA / birA <$> bioY-bioB*	*$ bioY2-X*
Clostridium botulinum #	**CB**	+	+	*[bioY-bioB-bioD*	*$ bioY*
Clostridium difficile #	**DF**	+	+	*$ bioB*	*$ bioY-yhfS-yhfT*
Clostridium perfringens	**CP**	+	+	*$ bioY-bioB-bioD / $ birA*	
Enterococcus faecalis	**EF**	+	+	none	*$ bioY*
Heliobacillus mobilis #	**HMO**	+	+	*[bioD / [bioA*	*$ bioY*
Listeria innocua	**LI**	+	+	none	*$ bioY*
Lactococcus lactis	**LL**	+	+	none	*birA-bioY <$> yhfT-yhfS*
Staphylococcus aureus	**SAX**	+	+	*$ bioD-bioA-bioB-bioF-bioW-bioX*	*$ bioY / $ yhfT-yhfS*
Streptococcus pneumoniae	**PN**	+	+	none	*$ bioY*
Streptococcus pyogenes	**ST**	+	+	none	*$ bioY / $ yhfS-yhfT*
Streptococcus equi #	**SEQ**	+	+	none?	*$ bioY / $ yhfS-yhfT*
Actinobacteriae					
Corynebacterium diphtheriae #	**DI**	0	+	*bioB1 / bioA-bioD / bioW-bioF / bioB2*	*bioY-cbiO-cbiQ*
Mycobacterium tuberculosis	**MT**	0	+	*bioB / bioA-bioF-bioD*	
Streptomyces coelicolor #	**SX**	0	+	*bioF <> bioB-bioA-bioD*	*bioY*
Thermomonospora fusca #	**TFU**	0	+	none?	*bioY-cbiO-cbiQ*
Others					
Aquifex aeolicus	**AA**	0	+	***X-X-bioB / bioW-X-X / X-X-bioD / bioA / bioC***	
Bacteroides fragilis #	**BX**	0	+	*bioA-bioF-bio(GC)-bioD*	
Chlamydia trachomatis	**QT**	0	+	*bioB-bioF-bioD-bioA-bioW*	*bioY*
Deinococcus radiodurans	**DR**	-	+	none	*bioY-cbiO-cbiQ*
Synechocystis sp.	**CY**	0	+	*bioB-bioY-lspA / bioD / bioF / bioA*	

To be continued

Prochlorococcus marinus	**CK**	0	+	*X-X-bioB / bioF-X-bioC-bioD-bioA*	*bioY-lspA*
Porphyromonas gingivalis #	**PG**	0	+	*bioB-bioA] / X-bioD / bioG-bioC / bioF]*	
1	2	3	4	5	6
Thermotoga maritima	**TM**	0	+	none	*fabH-X-fabK-bioY-fabD*
Thermus thermophilus #	**TQ**	+	+	*$ bioB*	*$ bioY*
Archaea					
Archaeoglobus fulgidus	**AG**	+	+	none	*$ bioY-cbiO-X*
Halobacterium sp.	**HSL**	+	+	none	*$ bioY-cbiO-X*
M. thermoautotrophicum	**TH**	0	+	none	*bioY*
Methanococcus jannaschii	**MJ**	0	+	*bioB1 / bioB2 <> bioW-bioF-bioD-bioA*	
Methanosarcina barkeri #, M. mazei	**MBA, MMZ**	+	+	none?	*$ bioY-cbiO-cbiO-cbiQ*
Pyrococcus abyssii, P. furiosus	**PO, PF**	+	+	none	*bioY <$> (D-b-birA)*
Pyrococcus horikoshii	**PH**	0	+	none	*bioY*

Finally, we dissected novel interesting examples of co-regulation of biotin-related genes using the positional analysis of biotin biosynthetic genes. We found positional linkage between *birA* and genes encoding biotin-dependent carboxylases in Actinobacteria and some archaea. Some of these genes are predicted to be regulated by biotin repressor. Several genomes have divergently transcribed *birA* and *bioY* genes with predicted BirA sites in their common regulatory region. Another example of co-regulation of *bioY* with genes of the fatty acid biosynthesis in *Thermotoga maritima* can be easily explained, as biotin is a required co-factor of carboxylase, the latter being involved in the first step of the fatty acid biosynthesis.

Discussion

Conservation of the BirA binding sites across large phylogenetic distances allows us to suggest that D-b-BirA is the first example of an ancient DNA-binding transcriptional factor common to eubacteria and archaea. In contrast, analysis of regulatory systems for biosynthesis of riboflavin and thiamin showed that they are operated by conserved RNA elements, the *RFN* element and the Thi-box, respectively.

A comparative analysis of the biotin regulon in complete genomes resulted in new functional predictions for the *bioY*, *yhfS* and *yhfT* genes. The first of them, *bioY*, widely distributed in eubacteria and archaea gene, is a member of the BirA regulon in all genomes containing D-b-BirA. Here we predict that *bioY* encodes a transporter for biotin or biotin-related compounds. Associated with BioY, the YhfS and YhfT proteins can be involved in conversion of this precursor compound to biotin. The systematic comparison of putative operon structures revealed the conserved gene string *bioY-cbiO-cbiQ* in some bacterial genomes. Such functional linkage between the putative ABC transporter CbiO-CbiQ and the biotin transporter BioY is enigmatic.

The enzymes mediating the first step of the biotin biosynthetic pathway are diverse. BioW and BioC represent two major types of enzymes involved in the synthesis of pimeloyl-CoA, a biotin precursor. We observed that various bacteria have different BioC-associated proteins (BioH, BioG, BioK, or BioZ). It can be explained either by utilization of different sources for biotin biosynthesis or by non-orthologous displacements of the BioC-linked proteins.

Acknowledgements

The authors are grateful to Andrei Osterman, Olga Vassieva and Alexandra Rachmaninova for helpful discussions. This study was partially supported by grants from INTAS (99-1476) and HHMI (55000309). It is a part of the "missing genes" project of Integrated Genomics.

REGULATION OF BACTERIAL RIBOFLAVIN GENES BY A CONSERVED RNA STRUCTURAL ELEMENT

Vitreschak A.G. [1]***, Rodionov D.A.*** [2*]***, Mironov A.A.*** [2,3]***, Gelfand M.S.*** [2,3]

[1] Institute for Problems of Information Transmission, Moscow, 101447, Russia
[2] State Scientific Center GosNIIGenetika, Moscow, 113545, Russia, e-mail: rodionov@genetika.ru
[3] Integrated Genomics–Moscow, P.O.Pox 348, Moscow, 117333, Russia
[*]Corresponding author

Key words: *genome analysis, riboflavin biosynthesis and transport, RFN element, bacteria*

Resume

We have identified riboflavin biosynthesis (RB) genes in almost all available bacterial genomes. In many diverse bacteria, the RB genes are regulated by a conserved RNA regulatory element *RFN*. Comparison of the nucleotide sequences around the *RFN* elements has revealed a set of conserved RNA secondary structures. In gram-positive bacteria, it includes the *RFN* element, terminator hairpin, and alternative antiterminator with the main stem overlapping both *RFN* and the terminator. In contrast, gram-negative bacteria have a sequestering hairpin that overlaps the Shine–Dalgarno (SD) sequence or the start codon of the first gene in the operon. Consequently, involvement of transcription and translation attenuation mechanisms in the regulation of the RB genes is proposed. Analysis of the operon structure shows that *RFN* predominantly regulates single RB genes in proteobacteria and the RB operon in most gram-positive bacteria. Moreover, single RB genes seem to be regulated at the level of translation, whereas the RBS operons are predicted to be regulated at the level of transcription. Analysis of the *RFN*-based regulation and operon structure allowed us to predict new riboflavin transporters, namely, *ypaA*, *impX,* and *pnuX* in the gram-positive bacteria and *rfnT* in rhizobia. Analysis of the *RFN* architecture, operon structure, and protein phylogeny identified several cases of likely horizontal transfer in *F. nucleatum* and two proteobacteria.

Introduction and Motivation

Riboflavin (vitamin B2) is an essential component of the basic metabolism because it is a precursor of the coenzymes flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN). Many microorganisms and all plants are able to synthesize riboflavin, but it is not produced by higher animals. The most well-studied system of riboflavin biosynthesis in bacteria is the *ribGBAH* operon of *Bacillus subtilis.* Furthermore, the riboflavin operons were studied in *Bacillus amyloliquefaciens* (Gusarov et al., 1997), *Actinobacillus pleuropneumoniae* (Fuller, Mulks, 1995), and *Photobacterium phosphoreum* (Lee et al., 1994). Recently, *B. subtilis* was shown to contain the riboflavin transport system YpaA (Kreneva et al., 2000). In contrast to *B. subtilis*, the riboflavin biosynthesis genes of *Escherichia coli* do not form a single operon but are scattered over the chromosome. Metabolic studies gave no evidence for any regulation of the riboflavin biosynthesis genes in *E. coli* (*E. coli Book*, ASM, 1994). On the other hand, flavin nucleotides, but not riboflavin, have an effector function for regulation of the riboflavin operon in *B. subtilis* (Lee et al., 2001). The regulatory region *ribO* located between the promoter and the coding region of the *ribGBAH* operon is involved in regulation. Recently, strong conservation of these regions in diverse bacteria was discovered. Moreover, a conserved RNA structure with five hairpins (the *RFN* element) corresponding to the *ribO* region was found to be involved in regulation of the riboflavin operon (Gelfand et al., 1999). However, the regulatory mechanism of riboflavin genes was not known. It is very interesting to find new *RFN* elements in available genomes and to analyze the regulation of riboflavin genes (see Results). Moreover, a possible regulatory mechanism of expression of riboflavin genes is suggested.

Results and Discussion

We have applied the RNA PATTERN program to scan all the available bacterial genomes for candidate *RFN* elements and identified the riboflavin biosynthesis genes in the listed bacterial genomes by similarity search. Totally, 61 *RFN* elements were found in 49 eubacterial genomes. All these elements are located upstream of the RB genes or potential riboflavin transport genes. Only spirochetes, mycoplasmas, and rickettsia have neither riboflavin genes nor *RFN* elements. The traditional RB gene names are different in *E. coli* and *B. subtilis,* and, for consistency, we use the *E. coli* gene names throughout. Thus, the *B. subtilis ribG*, *rib,* and *ribA* genes are renamed here to *ribD*, *ribE* and *ribBA*, respectively.
The riboflavin transporter gene *ypaA* was found in all the studied genomes of the *Bacillus/Clostridium* group except for *Bacillus halodurans*. Moreover, YpaA seems to be the only source of riboflavin in *Enterococcus faecalis* and *Streptococcus pyogenes*, as these genomes lack RB genes. Two more genomes containing *ypaA* are *Atopobium minutum* (actinomycete) and *Thermotoga maritima*. A *RFN* element precedes *ypaA* in the former, but not in the latter genome. Other actinomycetes seem to have a new, different riboflavin transporter. In *Thermomonospora fusca* and *Streptomyces coelicolor*, the RB operon consists of *ribE*, *RTFU01116* (named here *pnuX)*, *ribBA* and *ribH,* and has an upstream *RFN*

element. The *pnuX* gene is homologous to the nicotinamide mononucleotide transporter *pnuC* from enterobacteria and encodes a protein with six predicted transmembrane segments. Orthologs of the *pnuX* gene exist in two other actinomycetes, *Corynebacterium diphtheriae* and *Corynebacterium glutamicum*, and in the latter, *pnuX* is preceded by a *RFN* element. One more candidate riboflavin transporter, *impX*, is found in *Fusobacterium nucleatum* and *Desulfitobacterium halfniense,* with upstream *RFN* elements in both cases.

Most proteobacteria have some redundancy of the RB genes due to paralogs of the *ribH*, *ribBA* and *ribE* genes. Moreover, some genomes contain not only the fused *ribBA* gene, but also additional single *ribB* or *ribA* genes.

Phylogenetic analysis of the RB protein sequences detects possible horizontal transfer of the *ribDE(BA)H* operon from the *Bacillus/Clostridium* group to two Pasteurellaceae genomes, *Haemophilus ducreyi* and *Actinobacillus pleuropneumoniae.* The *RFN* elements upstream of this operon are of the non-gram-negative type.

Recently, it was shown that flavin mononucleotides (FMN) regulate expression of the RB operon in *B. subtilis* (Lee et al., 2001). We propose here a possible mechanism of the FMN-mediated regulation via the *RFN* element (Fig.). In general, two different types of regulation are suggested, the attenuation of transcription via antitermination mechanism and the attenuation of translation by sequestering of the Shine-Dalgarno box.

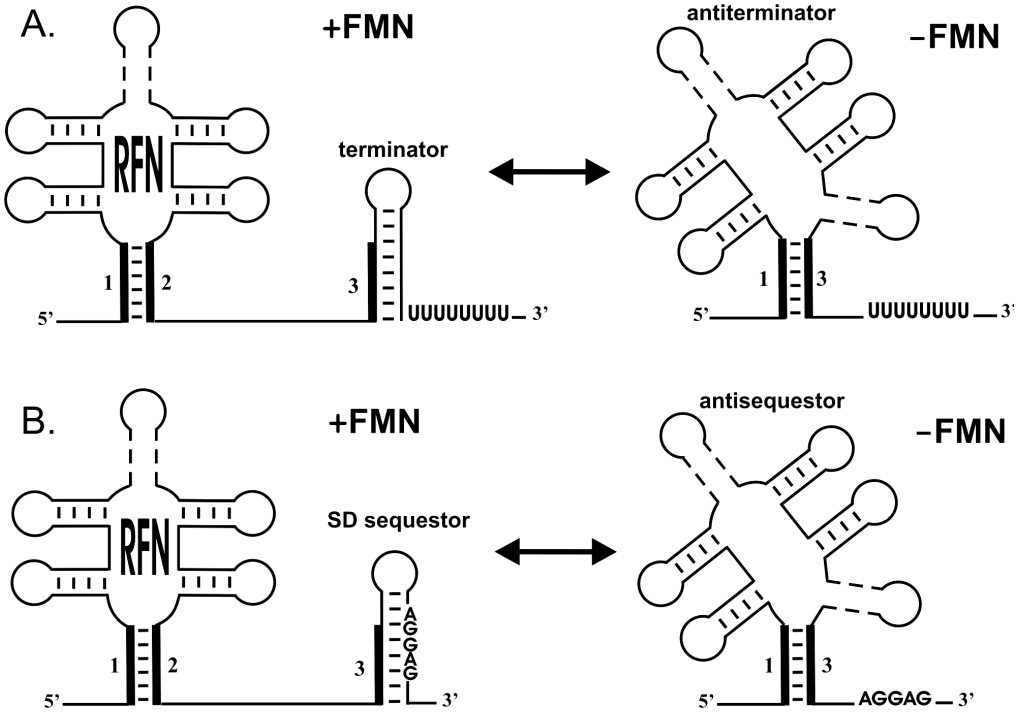


Fig. Predicted mechanism of the *RFN*-mediated regulation of riboflavin genes: (A) transcription attenuation and (B) translation attenuation.

In gram-positive bacteria, *Thermotoga maritima* and *Chloroflexus aurantiacus,* we have found terminator-like RNA structures located between the predicted *RFN* element and translational gene start of RB genes (Fig.). We found complementary fragments of RNA sequences that partially overlap both the first helix of *RFN* and the left stem of the terminator. Furthermore, these complementary fragments always form the first main helix of a more stable new alternative secondary structure with ΔG smaller than ΔG of the *RFN* element. We predict that this structure functions as an antiterminator, which is an alternative to both the *RFN* element and the terminator.

In other cases, mostly in gram-negative bacteria, the RNA hairpins downstream of the *RFN* element sequestor the ribosome-binding site (the Shine–Dalgarno box). In most cases, we have found a highly conserved sequence, GCCCTGA, which overlaps the proposed sequestor hairpin and is complementary to helix 1 of the *RFN* element. These two complementary sequences always form the stem of the RNA secondary structure, called here antisequestor, which is more stable than the *RFN* element. The proposed mechanism of translational regulation of the RB operons is similar to the termination–antitermination mechanism described above, but involves the SD-sequestor instead of the terminator.

Acknowledgements

We are grateful to Andrei Osterman, D.Perumov, and A.S.Mironov for useful discussions. This study was partially supported by grants from RFBR, HHMI, and INTAS.

References

1. Gusarov I.I., Kreneva R.A., Podcharniaev D.A., Iomantas I.V., Abalakina E.G., Stoinova N.V., Perumov D.A., Kozlov I.I. (1997). Riboflavin biosynthetic genes in *Bacillus amyloliquefaciens:* primary structure, organization, and regulation of activity. Mol. Biol. (Mosk.). 31:446-453.
2. Fuller T.E., Mulks M.H. (1995). Characterization of *Actinobacillus pleuropneumoniae* riboflavin biosynthesis genes. J. Bacteriol. 177:7265-7270.
3. Lee C.Y., O'Kane D.J., Meighen E.A. (1994). Riboflavin synthesis genes are linked with the lux operon of *Photobacterium phosphoreum.* J. Bacteriol. 176:2100-2104.
4. Kreneva R.A., Gelfand M.S., Mironov A.A., Iomantas I.A., Kozlov I.I., Mironov A.S., Perumov D.A. (2000). Study of the phenotypic occurrence of *ypaA* gene inactivation in *Bacillus subtilis*. Genetika**.** 36:1166-1168.
5. *E. coli.* Book, ASM, 1994.
6. Lee J.M., Zhang S., Saha S., Santa Anna S., Jiang C., Perkins J. (2001). RNA expression analysis using an antisense *Bacillus subtilis* genome array. J. Bacteriol. 183:7371-7380.
7. Gelfand M.S., Mironov A.A., Jomantas J., Kozlov Y.I., Perumov D.A. (1999). A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. Trends Genet. 15:439-442.

MCMC METHOD FOR IDENTIFICATION OF ALLELIC PATTERNS IN DATA WITH QUANTITATIVELY DESCRIBABLE PHENOTYPIC FEATURES

[1] Favorov A.V., [2] Ochs M.F.

[1] State Scientific Centre "GosNIIGenetica", 113545, Moscow, Russia, e-mail: favorov@sensi.org
[2] Fox Chase Cancer Center, 7701 Burholme Avenue, Philadelphia, PA, 19111, USA
*Corresponding author

Key words: *multigene traits, regulatory patterns extraction, Mann-Whitney-Wilcoxon distribution, Markov Chain Monte-Carlo, Metropolis-Hastings-Gibbs sampler*

Resume

Motivation: Some sophisticated phenotypic features (traits), e.g. susceptibility for multigene diseases level, are determined by complex interaction of a set of alleles for different loci. To investigate such an interaction, it can be extremely useful to split computationally the set of alleles into subsets (patterns), which influence the trait independently.
Results: A method for computational identification of such patterns from genetic and comparative phenotypic data is proposed. It utilizes Markov chain Monte-Carlo sampling of the space of all possible pattern sets. The function to be maximised is the Bayesian posterior of the proposition that all the patterns in the set are connected to the phenotypic feature (the pattern set's null-hypothesis rejection) given the data. The likelihood computation is based on a set of nonparametric (Wilcoxon-Mann-Whitney) tests. The algorithm was tested on a simple artificial test data set and identified the encoded patterns successfully.
Availability: The project is under development now, and the current version of the executables for FreeBSD and Windows NT console are available on E-Mail request.

Introduction

Complex interplay of a set of alleles for some collection of genes is believed to determine the level of some sophisticated phenotypic features (traits), e.g. susceptibility for multigene diseases. We suppose that it is possible to divide the set into some subsets of alleles that affect the trait independently. The allelic collections do not form anything like a linear space, so the concept of independency is not obvious. We treat some subsets of alleles (for different or for the same locus) as influencing a trait independently if the presence of each of these subsets in the genome affects the trait level when the influence of all the remaining subsets is removed. We refer to these subsets as allelic patterns, or just patterns. The interaction of genes associated with the feature inside the patterns still can be very complex. We refer to the patterns as promoting or suppressing, correspondingly to their effect, similar to predisposing or protective allele combinations in the case of disease susceptibility. The number of patterns must be determined independently at this point.
We formulate a numeric characteristic of quality (the null-hypothesis rejection posterior, see below) of a set of patterns given the genetic and phenotypic data. Then, we find the best set of patterns, which are combined from alleles from genetic data. For that purpose, we use a heuristic algorithm (a Metropolis-Hastings MCMC sampler, see below), which is able to explore all the possible sets of patterns and is drastically less time-consuming than complete enumeration.
The phenotypic feature levels are comparative characteristics. We avoid ascribing any sense to their relations besides "less", "equal" and "more" due to the difficulty of simple numerical description inherent in most measurements (e.g. disease level). We cannot remove analytically the influence of environmental factors on the manifestation of the genetic picture, but we postulate the positive correlation between the genetic factor and its phenotypic outcome degree. So, statistically individuals with more genetic predisposition will have higher levels of the feature. In essence, we treat the comparative level of a phenotypic trait as the level of the sum of genetic factors dominating the trait.
We denote the statement that a pattern's presence itself does not influence the trait level as the null hypothesis about this pattern. The null hypothesis about a pattern set is that at least one pattern's null hypothesis holds. So, the pattern set that has the maximal posterior of null-hypothesis rejection is the best split of the genetic data into a set of independently working allelic patterns.

Methods and Algorithms I: Posterior computation

Pattern sets and genetic data are involved in the posterior computation only together, as measures of presence or absence of every pattern in every genome. For example, consider the case of only one pattern in the set. All the phenotypic trait data is classified into two sets of levels with respect to presence or absence of the pattern in the individuals' genomes (we denote the sets cardinalities as g and h). The null hypothesis about the pattern states that the presence of the pattern does

not affect the trait level, so that the two sets of levels (classes) are random draws from the same distribution. To get the posterior related to this statement, we need to know the data likelihood given the null-hypothesis and the data likelihood given its rejection. We wish to ascribe only comparative meaning to the phenotypic trait levels, so we use criteria that concern pairwise level comparisons from the two classes. Every such comparison result is ascribed 1 or 0 corresponding to the direction of inequality that holds. The likelihood distribution for the sum of the comparisons results (n) given the null-hypothesis is the Wilcoxon-Mann-Whitney distribution with parameters g and h (see Van Der Varden, 1957; Statlib Algorithm AS 62, 1973; Bucchianico, 1996). We denote it $f(g,h,n)$. The null hypothesis negation means the presence of the pattern has an effect on the trait level. Our ignorance about the effect sign and strength leads to the likelihood distribution for n given null-hypothesis rejection that is flat between 0 and $g \cdot h$. As for priors, we ascribe a pre-given value to every pattern that is informative for our task. All other patterns, i.e. patterns which do not exist in the genome (or exist in every gene set), are to get null-hypothesis prior equal to 1. In addition to this posterior, the value of n gives us the sign of pattern's influence on the trait level, which can be positive, negative or undefined.

For *p* patterns, there are 2^p classes of trait levels instead of two, since each genome can carry every subset of patterns. For every pattern, we consider 2^{p-1} pairs of classes differ only by the pattern absence/presence. Then, we combine the results from the pairs obtaining the posterior of the null-hypothesis about the pattern and, in turn, combine the pattern posteriors to get the overall result. Finally, we obtain for the null-hypothesis rejection posterior given the data for the entire pattern set:

$$P\left(\overline{H_0} \mid data\right) = \prod_i \left[1 - \frac{\prod_e f\left(g_{ie}, h_{ie}, n_{ie}\right) \cdot P_i}{\prod_e f\left(g_{ie}, h_{ie}, n_{ie}\right) \cdot P_i + \left(\prod_e \frac{1}{g_{ie} \cdot h_{ie} + 1} \right) \left(1 - P_i\right)} \right] \tag{1}$$

The index i runs through all p patterns, while e counts the 2^{p-1} class pairs corresponding to the pattern. P_i denotes the null-hypothesis prior for the i-th pattern. If a pattern contains contradictory signs for different class pairs, we use its null-hypothesis prior P_i instead of its posterior (the ratio in the right term).

Methods and Algorithms II: The sampler

The algorithm used for the maximization of the null-hypothesis rejection posterior is a hybrid Metropolis-Hastings-Gibbs sampler (Robert, 1998; Besag et al., 1996). At each step it proposes a change in the current set of patterns, and, if the result has higher probability of the rejection of the null-hypothesis, accepts the change. Otherwise, it accepts the new variation with a probability proportional to the ratio of new and old posteriors. The collection of sets of patterns obtained by the Markov chain is distributed as it was sampled from the posterior distribution (1) (Robert, 1998; Besag et al., 1996), so the most visited set of patterns should be the most probable solution. The possible sequential changes are point change in an allele or a recombination of patterns. A point change can be an allele change, or an allele removal from or addition to a pattern.

Very often a pattern set is uninformative for given genetic data, i.e. at least one of its patterns does not divide all the genomes into two nonempty classes. The sampled posterior of an uninformative pattern set is equal to zero, hence the sampler would reject a transition from an informative set to an uninformative one, while a backwards transition would be accepted whatever.

Thus, a sampler, which is started from an uninformative pattern set, would wander a long time in a null-posterior region before reaching an informative one. To avoid this useless initial wandering, the program starts with a set of patterns, each carrying one allele whose presence in a gene set is correlated with the trait level. The sampling procedures starting from different sets for the same data were found to give almost the same result, so that the sampler appears reasonably unbiased.

Implementation and Results

The program implementing the algorithm was written in C and compiled using the gcc compiler from the GNU project. It was tested on simulated data that was created as follows. First, three patterns were created; the first contained 3 alleles, while two contained 2 alleles. Each pattern was assigned a role in the phenotypic trait, with two of the patterns being promoting, and one suppressing. For every pattern, patterns differing only in one allele (shadows) were created. Then, two copies of every pattern and three shadows were distributed randomly in 50 model gene sets, which were originally empty. All the empty loci positions were filled randomly. The only restriction placed on the distribution was that one locus could not contain more than two alleles. Then, a trait level was generated for each set, with a level equal to the sum of roles of all patterns contained in the gene set

Tests were made for 5 different data sets and for 5 starting points for each set. All the 25 starts have revealed the original set of original patterns as the most probable. Also, the algorithm has shown certain stability for noisy level data. When the Gaussian noise with σ equal to 1/5 of the weakest original pattern effect was added to the trait level data, the program still found correct pattern set.

Discussion

We have not included any domination model, implicitly assuming that every allele can affect the phenotypic feature regardless of its counterpart on the other chromosome. The algorithm could be modified to take into account a domination model, if desired.
The genomic typing data is preferred to be information about loci, which are known to be linked to the trait. The algorithm will not fail with the inclusion of uninformative loci, but any additional locus multiplies the computation time several fold.
The method does not depend on what kind of target trait we are interested in, although the trait must possess sensibly comparable characteristics (levels). As such, the algorithm can be applied to many types of data, possibly with minor algorithm changes being required.
The two main ideas presented here are the exploration of the space of all genetic pattern sets possibly tied to a phenotypic trait by a Metropolis-Hastings-Gibbs and the use of Wilcoxon nonparametric statistics to provide an association measure. Still the test we used to check the algorithm was quite artificial, it has shown that these ideas application can yield for the splitting of complex genetic data into independent regulatory patterns. The main restrictions of the proposal, which are revealed on this stage, are its execution time requirements and difficulties in interpretation of almost identical patterns interference.
In the future, we intend to improve the efficiency of the algorithm by using the posterior probabilities directly in the found pattern sets ranking. In addition, we are beginning tests on blinded simulation data, with the intention of applying the algorithm to real data in the near future.

Acknowledgements

This work was partially supported by the NIH Comprehensive Cancer Center Core grant CA06927. We are grateful to Giovanni Parmigiani for useful advise and for involvement in planning for future work.

References

1. Besag J., Green P., Higdon D., Mengersen K. (1996) Bayesian computation and Stochastic Systems. Statistical Sci. 10, 1, 3-66.
2. Bucchianico A. di (1996) Combinatorics, computer algebra and Wilcoxon-Mann-Whitney test. Memorandum COSOR 96-24, Eidhoven University of Technology.
3. Robert C.P. (1998) Discretization and MCMC Convergence Assessment, Springer-Verlag.
4. Statlib Algorithm AS 62 (1973) The distribution of the Mann-Whitney U-statistic. Applied Statistics. 22, 2.
5. Van Der Varden B.I. von (1957) Matematishe statistik. Springer-Verlag, Berlin-Gottingen-Heidelberg. ch. XII, section 63.

S-RNASES IN THREE PLANT FAMILIES WITH GAMETOPHYTIC SELF-INCOMPATIBILITY: PHYLOGENY RELATED TO A PUTATIVE NUMBER OF S-LOCI IN *ROSACEAE*

Alexeyenko A.V.

Breeding Modelling Section, Northern Caucasus Research Institute for Horticulture and Viticulture, Krasnodar, 350901, Russia, e-mail: avalex99@mail.ru

Key words: *self-incompatibility, higher plants, Rosaceae, consensus sequence, multiple alignment, multiple alleles*

Resume

Motivation: Nearly all the molecular investigations to date reported deal with single locus model of self-incompatibility (SI) in *Rosaceae* and, respectively, find one product (*S*-glycoprotein) of the *S*-gene per gaplotype. Yet, it does not correspond to the abundant experimental data in a number of species. An idea of the investigation is to find a by-pass prove for existence of paralogous loci in at least some genera of *Rosaceae* that still escape direct detection. The problem was complicated by highest degree of allele sequence variability of a population due to specific function of the genes studied.
Results: Isolated phylogenetic position of the known S-loci in two *Rosaceae* subfamilies has been shown. Use of the Corpet's consensus amino acid sequences and AliBee multiple alignment proved to be effective to operate with objects of higher intra-species and lower inter-species variability.
Introduction: SI is an important feature of higher plants to prevent fertilization with their own pollen and, thus, to force outcrossing. Gametophytic SI (determined by a gaplotype of the pollen grain rather than whole genotype of the pollinator plant that is the case of sporophytic SI) occurs, e.g., in the species of *Rosaceae*, *Veronicaceae* and *Solanaceae* families. The pollen grains carrying the same allele as one of the two of the pollinated plant have to be recognized and rejected during pollen tube growth. The pistil products of *S*-gene, known as glycoproteins, are characterized as RNases of *T2* class and preserve specific RNase activity (while the nature and function of the pollen-side product remain undiscovered in these families).
Being switched off in many species (presumably well adapted to the environment conditions) SI persists in most others. Anyway, respective nucleotide sequences can be found in almost every higher plant genome and are the objects of the molecular genetic study. The specificity of *S*-gene function results in a great number of the alleles and highest heterozygocity in the most part of populations studied. Plenty of alleles of the respective *S*-gene have been found and well characterized in these families. Meanwhile, there is no consideration found of possible existence of other loci controlling rejection of self-pollen. Yet, our study (Alexeyenko, Volchkov, 1999) proved that at least three gametophytic genes control SI in European population of the domestic pear *Pyrus communis.* Taking into consideration that SI loci are highly evolutionary conserved (excluding variability of specific *S*-glycoprotein domains causing higher polymorphism in the population), there is a serious doubt that different system exists in close species such as apple *Malus domestica* and Japanese pear *Pyrus pyrifolia* (both belong to *Maloideae* subfamily). That is partially confirmed by various observations in apple witnessing for a pattern of pollen tube segregation similar to pear (Doutova, Broothaerts; personal communications) while every new allele is searched and attributed to the single S-locus.
Dozens of *S*-gene sequences have been submitted to the GenBank for these two species as well as for *Amygdaloideae* (another *Rosaceae* subfamily) and for *Solanaceae* and *Veronicaceae* species. At the same time, there are no entries for *Pyrus communis.* Because the field genetic experiments in tree plants are difficult and time consuming, it is worthwhile to employ the techniques of computational biology to try to solve at least a part of the problem.

Methods and Algorithms

All the sequences under investigation have been obtained as proteins from Genbank using keywords "*Family* AND (S-RNase OR S-gene OR S-like OR S-glycoprotein)" where *Family* is *Rosaceae*, *Veronicaceae* or *Solanaceae*, respectively. Then, every sample was inspected carefully and every duplicate, incomplete or irrelevant sequence has been deleted. Consensus sequences has been derived for the genera rather than species because of lack of the data available (some genera are represented by just one species) and high conservation of the self-incompatibility systems (see below). The only exclusion was made for three *Amygdaloideae* species because of importance of their data (to this, the three species were considered as separate genera not long ago and have been included in *Prunus* rather artificially). To produce consensae, MultAlin web suite was used (Corpet, 1989) with BLOSUM62 matrix. Refined consensae without gaps were computed at the second stage of the process when Alignment and tree description (rfd) option was selected. Letter coloring denoted degree of the position conservation through the alignment and was used to mask segments of

hypervariability when necessary. While doing so, we masked letters with consensus less than 70%. Single letters were not masked when they occurred between two segments of high consensus, and vice versa. Consensus sequences were used as taxa representatives to produce matrices of pairwise dissimilarities between the sequences while multiple alignments provided by GeneBee-NET service (Brodsky et al., 1995) using AliBee program with Dayhoff matrix. Although AliBee program provides tree clusterization, it was decided to use specialized software to select a proper method. The dissimilarity matrices from AliBee were used in *Statistica* package (http://www.statsoft.com) for Ward's clusterization (StatSoft, Inc., 1999).

Implementation and Results

S-glycoproteins of *Rosaceae* have been reported to consist of a number of conserved domains intermitted with hypervariable regions. Glycan chains are attached mostly to the last ones and are thought to determine the allele specificity in SI reaction (Ishimizu et al., 1999). Presence of the hypervariable parts in a gene makes inferring phylogeny rather specific. In fact, there are two kinds of variability; those resulted from population polymorphism have nothing in common with usual phylogenetic and evolution relationships and shall be excluded. The problem was solved with producing multiple alignments using reliable and representative *S*-allele sequences of, usually, one or two species per taxon. Totally, 10 consensus representatives were produced (Table).

Table. Taxon representative sequences and their sources.

Family (subfamily)	Representative	Entries used	Referred to as
Rosaceae (Maloideae)	*Malus domestica*	12	Malus
	Pyrus pyrifolia	7	Pyrus
Rosaceae (Amygdaloideae)	*Prunus avium*	6	P. avium
	P. dulcis	11	P. dulcis
	P. mume	10	P. mume
Solanaceae	*Lycopersicon peruvianum*	18	Lycopersicon
	Nicotiana alata	11	Nicotiana
	Solanum chacoense + S. tuberosum	12	Solanum
	Petunia integrifolia	7	Petunia
Veronicaceae	*Antirrhinum hispanicum*	3	Antirrhinum

High conservation of the gametophytic S-genes (apart from allele polymorphism) is well known and reported elsewhere (Broothaerts et al., 1995; Sassa et al., 1996; Uyenoyama, 1995). So, on condition of single S-gene control throughout the family it would not be astonishing to find two subfamilies of *Rosaceae* well similar. To learn it, we employed clusterization of the 10 sequences using dissimilarity matrix, first without masking hypervariable regions. And, as one can see (Fig.), amalgamation of the genera that belong to a subfamily occurs at the steps 1-3 of the clusterization while the two subfamilies amalgamate with each other at the last step. After masking, no essential changes were observed in topology and amalgamation scheduling (not shown). In other words, the difference is greater than between any of them and another family studied.

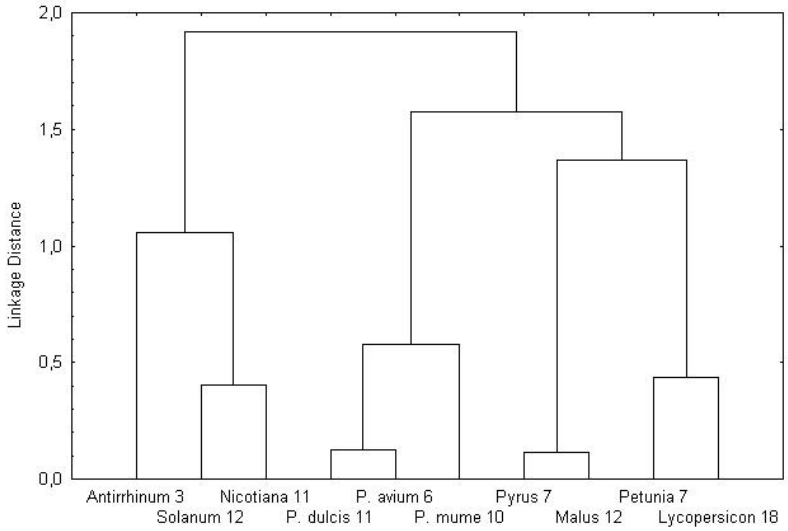


Fig. Cluster tree (Ward's method) of the representative S-protein sequences similarity in 10 higher plant genera with gametophytic self-incompatibility (without masking). *Numbers* and *names*, see Table.

To check whether the comparison of alignments inferred from masked and not masked sequences is redundant or not, the following test was performed. Control random segments of a 20-30 amino acid residues were included into a *Rosaceae* sequence and then used to produce a new dissimilarity matrix and a cluster tree. Almost every such invasion resulted in strong distortion of not only amalgamation schedule but also the tree topology itself (not shown).

Discussion

The selected algorithms proved to be sensitive enough to detect sequence relationships. At the same time, they are robust to occasional disorders in sequence reading and handling as well as to occurred hypervariability. The distinction between *S*-alleles in *Maloideae* and *Amigdaloideae* goes far beyond a usual shift with molecular clock and can be explained with functional reason or genome duplication. By our opinion, this can be a witness in favor of existence of another locus (loci) in the *Rosaceae* species. This hypothesis is partially confirmed by the recognized fact of polyploidy origin of the apple subfamily having $2n=34$ (while *Amigdaloideae* have $2n=16$). Along with known *S*-loci another one in *Maloideae* can exist that are closer to the *Amigdaloideae* pattern and have to be found in near future along with mysterious pollen product of *S*-gene.

References

1. Alexeyenko A.V., Volchkov Yu.A. (1999) Elucidation of genetic system for self-incompatibility in pear by means of genetic and mathematic modelling Selskokhozyaystvennaya biologia (Agricultural biology). 5, 95-102 (Russ.).
2. Brodsky L.I., Ivanov V.V., Kalaidzidis Ya.L., Leontovich A.M., Nikolaev V.K., Feranchuk S.I., Drachev V.A. (1995) GeneBee-NET: Internet-based server for analyzing biopolymers structure Biochemistry. 60, 8, 923-928.
3. Broothaerts W., Janssens G., Proost P., Broekaert W. (1995) cDNA cloning and molecular analysis of two self-incompatibility alleles from apple. Plant Mol. Biol. 27, 449-511.
4. Corpet F. (1989) Multiple sequence alignment with hierarchical clustering. Nucl. Acids Res. 16 (22), 10881-10890.
5. Ishimizu T., Mitsukami Y., Miyagi M., Shinkawa T., Natsuka S., Hase S., Sakiyama F., Norioka S. (1999) Presence of asparagine-linked N-acetylglucosamine and chitobiose in Pyrus pyrifolia S-RNases associated with gametophytic self-incompatibility. Eur. J. Biochem. 263, 624-634.
6. StatSoft Inc. (1999). STATISTICA for Windows [Computer program manual]. Tulsa, OK.
7. Uyenoyama M. K. (1995) On the evolution of genetic incompatibility systems. VI. A three-locus modifier model for the origin of gametophytic self-incompatibility. Genetics. 128, 453-469.

SURVEY OF HUMAN NON-SYNONYMOUS SNPs

Ramensky V.E. [1*]***, Bork P.*** [2,3]***, Sunyaev S.R.*** [1,2]

[1] Engelhardt Institute of Molecular Biology (EIMB), 119991 Moscow, Russia, e-mail: ramensky@imb.ac.ru
[2] European Molecular Biology Laboratory (EMBL), Meyerhofstr. 1, 69117 Heidelberg, Germany
[3] Max-Delbrueck Center for Molecular Medicine (MDC), Robert-Roessle-Strasse 10, 13122 Berlin, Germany
[*] corresponding author: e-mail: ramensky@imb.ac.ru

Key words: *SNP, complex disease, genetic variation, amino acid substitution, protein structure, protein function, comparative genomics*

Resume

Motivation: Single nucleotide polymorphisms (SNPs) comprise the bulk of human genetic variation. One of the main goals of SNP research is to understand genetics of the human phenotype variation, especially from the complex diseases perspective. Non-synonymous coding SNPs (nsSNPs) are a group of SNPs, which together with SNPs in regulatory regions are believed to have the highest impact on phenotype.
Results: We present PolyPhen, a WWW tool for prediction of the effect of an nsSNP on protein structure and function. The developed method enabled the comprehensive analysis of all human nsSNPs currently available via HGVBase, a database of human genetic variation. It was shown that the selection pressure against deleterious SNPs depends on molecular function of the protein, although shows no correlation with several other protein features considered. The strongest selective constraint was detected for proteins involved in transcription regulation.
Availability: The PolyPhen server and data collection are available at {{http://www.bork.embl-heidelberg.de/PolyPhen}}.

Introduction

Many of human SNPs are believed to cause phenotypic differences between human individuals. However, identifying SNPs responsible for specific phenotypes appears to be a difficult problem. Association studies (Risch, Merikangas, 1996) and candidate gene studies (Risch, 2000; Emahazion et al., 2001) have been proposed as experimental techniques to identify SNPs underlying complex, mostly disease, phenotypes. Both approaches face the problem of testing the overwhelming number of candidate SNPs. A possible way to overcome this problem is to use the bioinformatics expertise to discriminate between neutral SNPs, which constitute majority of genetic variation, and SNPs of likely functional importance. We focus on non-synonymous SNPs (nsSNPs), i.e. SNPs located in coding regions and resulting in amino acid variation in protein products. It was shown in several recent studies (Sunyaev et al., 2000; Sunyaev et al., 2001; Wang, Moult, 2001; Chasman, Adams, 2001; Ng, Henikoff, 2001; Ferrer-Costa et al., 2002) that impact of amino acid allelic variants on protein structure and function could be predicted via analysis of multiple sequence alignments and protein 3D-structures. As we demonstrated in an earlier work, these predictions correlate with the effect of natural selection seen as an excess of rare alleles (Sunyaev et al., 2001). Therefore, predictions at the molecular level reveal SNPs affecting actual phenotypes.

Methods and Algorithms

PolyPhen (=*Poly*morphism *Phen*otyping) is an automatic WWW tool (Fig.) for prediction of possible impact of an amino acid substitution on the structure and function of a human protein. This prediction is based on straightforward empirical rules, which are applied to the sequence, phylogenetic and structural information characterizing the substitution. PolyPhen input is amino acid sequence of a protein or the database ID or accession number together with sequence position and two amino acid variants characterizing polymorphism. For a given amino acid substitution in a human protein, PolyPhen performs several steps:
1) Sequence-based characterization of the substitution site. PolyPhen uses Feature Table section of the corresponding SWALL database (Apweiler 2000) entry (if available) and checks whether the amino acid replacement occurs at a site which is annotated as DISULFID, THIOLEST, THIOETH bond or BINDING, ACT_SITE, SITE, etc., site.
2) Calculation of multiple alignment-derived PSIC profile scores for two amino acid variants. An amino acid replacement may be incompatible with the spectrum of substitutions observed at the position in the family of homologous proteins. PolyPhen identifies homologues of the input sequences via BLAST search (Altschul et al., 1990) in the non-redundant protein sequence database. The resulting multiple alignment is used by the PSIC algorithm (Position-Specific Independent Counts, Sunyaev et al., 1999) to calculate the so-called profile matrix. Elements of the matrix (profile scores) are logarithmic ratios of the likelihood of given amino acid occurring at a particular position to the likelihood of this amino acid occurring at any position (background frequency).
3) Calculation of structural parameters and contacts. PolyPhen BLASTs query sequence against protein structure database and maps the substitution position onto the corresponding positions in homologous proteins with known structure. The conservation of structural characteristics in proteins with considerable (>50%) homology allows one to use

the homologs when the spatial structure of a protein under study is not known. PolyPhen uses DSSP database (Kabsch, Sander, 1983) to obtain the following structural parameters for the mapped amino acid residues: (a) secondary structure, (b) solvent accessible surface area, and (c) phi-psi dihedral angles. The following values are calculated by PolyPhen: (d) normed accessible surface area, (e) change in accessible surface propensity, or "hydrophobic potential", resulting from the substitution, (f) change in residue side chain volume, (g) region of the phi-psi map derived from the residue dihedral angles, (h) normalized B-factor (temperature factor) for the residue. Since the presence of specific spatial contacts of a residue may reveal its role for the protein function, PolyPhen checks three types of contacts for a variable amino acid residue: (i) contacts with ligands, (ii) interchain contacts, and (iii) contacts with functional sites where the location of the latter is taken from the SWALL Feature Table.

The retrieved and calculated data characterizing the substitution are used in decision rules which predict that a nsSNP is

- *probably damaging*, i.e., it is with high confidence supposed to affect protein function or structure,
- *possibly damaging*, i.e., it is supposed to affect protein function or structure,
- *benign*, most likely lacking any phenotypic effect,
- *unknown*, when in some rare cases, the lack of data do not allow PolyPhen to make a prediction.

The rules successfully predict ~82% of disease-causing mutations annotated in the Swiss-Prot database (Apweiler, 2000) and produce about ~8% of false-positives in the control set of between-species substitutions. Multiple alignment-based profile scores provided major contribution to the prediction, making predictions reasonably reliable even in cases of proteins with no homologue with known 3D structure.

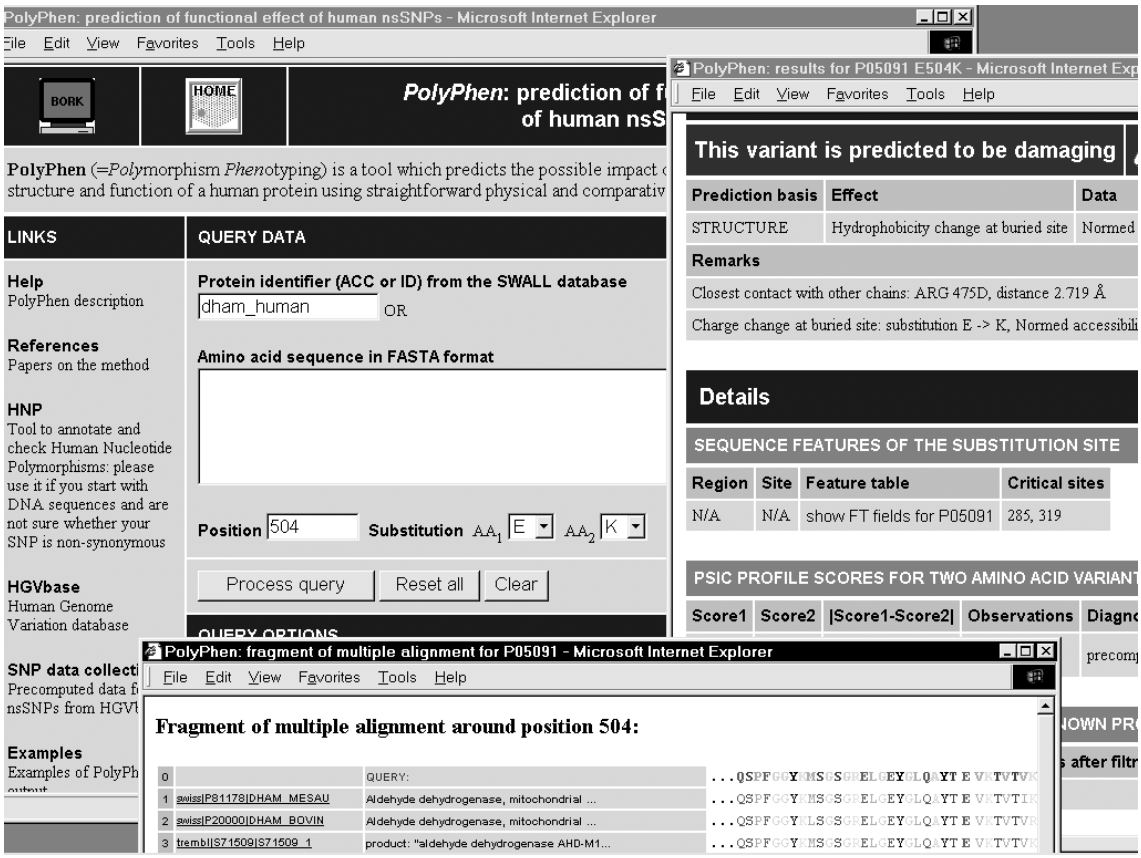


Fig. PolyPhen is an automatic tool for analysis of impact of human nsSNPs upon protein structure and function.

Implementation and Results

The server was used to annotate all SNPs deposited in the HGVbase (Fredman et al., 2001) database (Version 12). Collection of annotated 11,152 nsSNPs is available at {{http://www.bork.embl-heidelberg.de/PolyPhen/data}}. PolyPhen analysis was possible for 9,165 (82%) of these nsSNPs because the remaining ones have been mapped to proteins with no reasonably close homologous sequences available in the SWALL database for multiple alignment or structural analysis. Of these nsSNPs, 6,317 (69%) were predicted as benign, whereas 2,848 (31%) as damaging. Since only 1,026 nsSNPs were mapped to proteins with at least 50% sequence identity to a protein with known 3D-structure, the analysis for the most part of nsSNPs was performed on the basis of multiple alignment and sequence information. For proteins with available structural characteristics, parameters of hydrophobic core stability (e.g., change of volume or hydrophobic potential in the hydrophobic core) are better predictors as compared to those responsible for functionality (e.g., distance to

ligand). This supports the view that most of disease mutations and supposedly deleterious nsSNPs affect protein stability rather than functionality (Wang, Moult, 2001).

We used the GO (The Gene Ontology Consortium 2001) and SCOP (Lo Conte et al., 2002) classifications to divide all proteins with SNPs into large classes according to their (i) secondary structure class, (ii) biological process, (iii) localization, and (iv) molecular function. Contrary to our expectations we did not detect a significant correlation of the selective pressure against deleterious nsSNPs for secondary structure class, localization and biological process. In contrast, molecular function of the protein showed statistically significant association with the strength of selective pressure. The functional class showing the highest selective pressure against deleterious nsSNPs is the class of transcription factors. Enzymes are the class of proteins with the lowest selective pressure.

Discussion

SNPs involved in human complex phenotypes do not necessary determine the phenotype. Their effect depends on many other genetic and environmental components. In other words, SNPs may comprise risk factors of getting specific phenotypes in the statistical sense. Therefore the effect of a particular SNP on the phenotype might be seen only as frequency difference between individuals that display the phenotype and unaffected controls. The PolyPhen can be used to evaluate if the association under question can indeed have the functional meaning. Besides that, the data collection of nsSNPs already annotated by the PolyPhen provides a source of functionally annotated nsSNPs. The collection might be a useful resource for selection of nsSNPs for candidate gene based association studies.

Acknowledgements

Authors are thankful to Evgenia Kriventseva and Alexey Kondrashov for help and discussions.

References

1. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-410.
2. Apweiler R. (2000) Protein sequence databases. Adv. Protein Chem. 54, 31-71.
3. Chasman D., Adams R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J. Mol. Biol. 307, 683-706.
4. Emahazion T., Feuk L., Jobs M., Sawyer S.L., Fredman D. St., Clair D., Prince J.A., Brookes A.J. (2001) SNP association studies in Alzheimer's disease highlight problem for complex disease analysis. Trends Genet. 17, 407-413.
5. Ferrer-Costa C., Orozco M., de la Cruz X. (2002) Characterization of Disease-associated Single Amino Acid Polymorphisms in Terms of Sequence and Structure Properties. J. Mol. Biol. 315, 771-786.
6. Fredman D., Siegfried M., Yuan Y.P., Bork P., Lehvaslaiho H., Brookes A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucl. Acids Res. 30, 387-391.
7. The Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. Genome Res. 11, 1425-1433.
8. Kabsch W., Sander C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577-2637.
9. Lo Conte L., Brenner S.E., Hubbard T.J., Chothia C., Murzin A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. Nucl. Acids Res. 30, 264-267.
10. Ng P.C., Henikoff S. (2001) Predicting deleterious amino acid substitutions. Genome Res. 11, 863-874.
11. Risch N., Merikangas K. (1996) The future of genetic studies of complex human diseases. Science. 273, 1516-1517.
12. Risch N.J. (2000) Searching for genetic determinants in the new millennium. Nature. 15, 847-856.
13. Sunyaev S., Ramensky V., Bork P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. Trends Genet. 16, 198-200.
14. Sunyaev S., Ramensky V., Koch I., Lathe W. 3rd, Kondrashov A.S., Bork P. (2001) Prediction of deleterious human alleles. Hum. Mol. Genet. 10, 591-597.
15. Sunyaev S.R., Eisenhaber F., Rodchenkov I.V., Eisenhaber B., Tumanyan V.G., Kuznetsov E.N. (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng. 12, 387-394
16. Wang Z., Moult J. (2001) SNPs, protein structure, and disease. Hum. Mutat. 17, 263-270.

COMPUTATIONAL BIOLOGY AND ANALYSIS OF HUMAN POPULATIONS WITH USE OF DNA MARKERS

Zhivotovsky L.A.

Institute of General Genetics, Moscow, Russia, e-mail: lev@vigg.ru

Key words: *humans, DNA marker, evolution, population, microsatellite, STR, SNP*

Resume

I consider approaches to inferring information from data on DNA markers (SNPs and STRs) in human populations based on evolutionary models and corresponding statistical methods.

The subject

Large genetic data become available from worldwide human populations following the achievements in DNA technology. In particular, these data include many hundreds (and soon thousands) of DNA markers, such as SNPs (single nucleotide polymorphisms) and STRs (short tandem repeat polymorphisms) dispersed over the genome. Their distribution among individuals follows a complex probabilistic law which, in turn, is determined by the past population events, in particular by ancient population dynamics, migration flow and divergence, which are usually are not known and/or cannot be directly dated. To reveal information on these from DNA data the tools of mathematics and statistics need to be combined. In the talk, I shall concern with two kinds of problems that necessitate such combined tools for application to genetic data: 1) dating ancient demographic events; 2) analysis of associations of DNA markers and phenotypic traits.

Analysis of ancient evolutionary events is mainly based on distribution of DNA markers in the present populations. To interpret their distribution patterns in terms of population divergence, growth, migration flow, etc., dynamic models has to be analyzed. This involves a theory of stochastic processes and analysis of the higher probabilistic moments. The equations can be solved analytically and then further analyzed numerically or can be investigated via computer simulations. Such a model includes parameters for mutation rate, population size, rate of gene flow and population growth, as well as statistical estimators for divergence time and population expansion. The latter can be statistically estimated from population DNA data. In the talk I apply this approach to constructing phylogenetic population trees for modern humans.

Associations between DNA markers and phenotypic traits determined by a gene with an unknown chromosomal location is an important application of DNA technology to human genetics, especially in cases of marker-disease associations to map damaged genes. This approach is based on analysis of linkage disequilibrium in a population between a set of many tightly linked markers with known positions on the chromosome and the disease phenotype. It plots p-values (that is, the significance levels for the corresponding linkage disequilibria) against markers. Then it searches for the markers with maximal p-values, the latter are assumed to be physically close to the damaged gene. Sequencing in the neighborhood of the markers is the further step in identification of the gene, it can be successful only if the markers are those close to the gene. However, this approach meets some statistical difficulties because sample sizes are usually not large, and the markers with large p-values may not be tightly linked to the gene. Also, the usual statistical procedures assume statistical independence of p-values obtained in such studies, which is not always the case because evolutionarly introduced stochasticity comes up with linkage disequilibrium between arbitrary loci by chance. Therefore, more sophisticated statistical tools should be developed and applied to genetic database that carry information on hundreds and thousand markers (so, hundreds and thousand p-values). Statistical approaches to solving this problem are considered in the talk. The talk will be based on recently obtained data and on the following published material.

References

1. Knight A., Underhill P.A., Zhivotovsky L.A., Ruhlen M., Mountain J.L. (2002) African Y chromosome and mtDNA evidence suggests that all living humans descend from speakers of a click language. Proc. Natl Acad. Sci. USA. (subm.).
2. Zaykin D.V., Zhivotovsky L.A., Westfall P.H., Weir B.S. (2002) Truncated product method for combining p-values. Genet. Epidem. 22: 170-185.
3. Zhivotovsky L.A., Goldstein D.B., Feldman M.W. (2001) Genetic sampling error of distance $(\delta\mu)^2$ and variation in mutation rate among microsatellite loci. Mol. Biol. Evol. 18: 2141-2145.
4. Zhivotovsky LA. (2001) Estimating divergence time with use of microsatellite genetic distances: impacts of population growth and gene flow. Mol. Biol. Evol. 18: 700-709.
5. Zhivotovsky L.A., Ahmed S., Wang W., Bittles A.H. (2001) The forensic DNA implications of genetic differentiation between endogamous communities. Forensic Sci. Intern. 119: 269-272.
6. Zhivotovsky L.A., Bennett L., Bowcock A.M., Feldman M.W. (2000) Human population expansion and microsatellite variation. Mol. Biol. Evol. 17: 757-767.

7. Jin L., Baskett M.L., Cavalli-Sforza L.L., Zhivotovsky L.A., Feldman M.W., Rosenberg N.A. (2000) Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. Ann. of Hum. Genet. 64: 117-134.

THE CHANNEL CAPACITY OF SELECTIVE BREEDING: ULTIMATE LIMITS ON THE AMOUNT OF INFORMATION MAINTAINABLE IN THE GENOME

Watkins C.J.C.H.

Department of Computer Science, Royal Holloway, University of London,
Egham Hill, Egham, Surrey TW20 0EX, United Kingdom, e-mail: C.Watkins@cs.rhul.ac.uk

Key words: *evolution, information theory, channel capacity, selective breeding, genotype, sexual reproduction, asexual reproduction, genetic architecture*

Resume

Motivation: Genomes contain the information for constructing organisms. In some sense, this information is put in by selection, and it is degraded by mutation and genetic drift. It is natural to pose some basic questions of principle. What is the maximum amount of information that can be maintained at mutation-selection equilibrium? Can organisms in principle become arbitrarily complex, or, for a given mutation rate and intensity of selection, is there a limit to the amount of information that can even in principle be maintained in the genome? What method of encoding information in the genome allows the greatest amount of information to be maintained for the least selective cost? Indeed, how can the "information from selection" even in principle be defined? As far as we are aware, these questions have not previously been satisfactorily posed nor answered.
Results: We make these questions precise and answer them by modelling selective breeding as a communication channel. We show how the information-theoretic capacity of this channel provides a measure of the amount of information that can be put into or maintained in the genome by selection. The channel capacity is computed for some simple genetic models.
A striking result is that for a large population in mutation-selection equilibrium, the amount of information that can be maintained using a diffuse encoding analogous to an error-correcting code is vastly greater than the amount of information that can be maintained if information is encoded as exact sequences of nucleotides.

Introduction

Organisms are shaped by selective breeding, where the selection may be natural or artificial. In some sense, selective breeding introduces and maintains the large amounts of information necessary to construct complex organisms. It is natural to ask some basic questions about the total amount of information that could be maintained in the genomes of a species through selection, whether natural or artificial.
First, how can the information that is the result of selection even in principle be defined? In real populations of members of a species, there are many common features of the genomes that are accidendal and of no adaptive significance. How can we define, even in principle, the amount of information in the genomes of a species that is the result of selection?
Next, the amount of information that may be maintained will depend upon

- The mode of reproduction (sexual or asexual).
- The mutation rate per locus per generation.
- The intensity of selection (very intense selection may maintain more information).
- The method of encoding of the information in the genome.

In classical population genetics, such as in the textbook of Crow and Kimura (1970), the influence of mutations on an organism is described in terms of to what degree the mutation is advantageous or deleterious. However, there is a different and complementary view: how does selection affect the amount of information in the genome as a whole? Some mutations become fixed, others are lost: in a large genome in which many mutations occur, selection influences the fate of mutations statistically, but will not determine the fate of all mutations that occur. In slightly influencing the fates of many mutations, how does selection influence the amount of information stored in the genome as a whole?
In particular, it has been rediscovered many times that sexual reproduction eliminates deleterious mutations far more efficiently than asexual reproduction. In genetics, Crow and Kimura (1979) is an early paper, while Kondrashov (1988) provides a review: in computer science, independent analyses are Muhlenbein (1993), Baum (1995), and Mackay (1999).
We give an alternative informational analysis, and we show that the maximum maintainable amount of information in the genome is much larger with sexual than with asexual reproduction, and that selection can be more effective when applied to large numbers of loci each with small effect than when it is applied to a small number of loci, each with large effect.

Model

Selective breeding as a communication channel

Very briefly, we regard selective breeding as a communication channel in which the "message sent" is the rule for selecting which organisms in each generation to breed from, and the "message received" is a single organism sampled once mutation-selection equilibrium is reached. The information transmissible from selection rule to final organism is a measure of the maximum extent to which the genome of the organism could be influenced by selective breeding: it is a measure of the "adaptive capacity" of the organisms. The "message received" is defined to a single organism because we are interested in the information from selection that is present in the entire final breeding population. For example, the genetic information that makes a poodle a poodle must be present in almost all of a breeding population of poodles.

A simple genetic model

We use a simplified model, in which genomes are fixed-length vectors with elements that are 0 or 1. We assume that there is a large population in which all loci are in full linkage equilibrium at all times, so that for a randomly sampled genome, the values at all loci are statistically independent. (Approximate results for small populations may be obtained using the standard diffusion approximations given in Crow and Kimura (1970)).

Let the length of all genomes be L. Let the mutation rate U be defined as the fraction of loci per generation that change from 0 to 1, or from 1 to 0: mutations in each direction are assumed to be equally likely. All mutations are assumed to be point mutations that occur independently at all loci: no mutations that are insertions or deletions occur.

The exact form of the constraint on intensity is not important: we will assume there is truncation selection in which the fittest 50% of genomes are selected to breed the next generation. Subject to this constraint, it turns out that maximum channel capacity is achieved with selection rules of the form: for some ideal genome **g**, select the 50% of genomes in the population that agree in most positions (are closer in Hamming distance) to **g**. There are 2^L possible "ideal" genomes **g**, and hence 2^L possible selection rules of this form. Two extreme forms of genetic encoding are easy to analyse:

1. Encoding information as an exact nucleotide sequence

Suppose that we require the genomes of all members of the population to be, with high probability, exactly equal to **g**. For small U, the number of mutations in a child-genome will be Poisson distributed with mean UL. For truncation selection to restore the population to purity, the fraction of offspring with no mutations must be greater than ½. We require therefore that $e^{-LU} \geq \frac{1}{2}$, which implies that $L \leq \frac{\ln 2}{U}$. In this selection regime, each **g** yields a distinct equilibrium population, so for given U and optimising over L, the channel capacity is $\frac{\ln 2}{U}$ bits.

2. Encoding information as a long but highly variable nucleotide sequence

Assume that the population is large. For large L, in equilibrium the population will be polymorphic, and a randomly drawn genome will not agree with **g** at all loci. Let a locus at which a genome agrees with **g** be termed an *agreement*. Let the mean fraction of agreements in genomes drawn from the equilibrium population be p, where $1 > p > ½$. We calculate p as follows. First, note that for a large enough population, at *all* loci the fraction of alleles identical to the corresponding allele in **g** will be close to p. The variance of the fraction of agreements in individual genomes drawn from the population is therefore $\sigma^2 = \frac{p(1-p)}{L}$. In each generation, 50% truncation selection increases p by $\frac{\sigma}{\sqrt{2\pi}}$, and mutation reduces p towards ½ by $2U(p - ½)$. The equilibrium equation is therefore

$2U(p-\frac{1}{2}) = \sqrt{\frac{p(1-p)}{2\pi L}}$. Solving for p we obtain $p-\frac{1}{2} = \frac{1}{2\sqrt{8\pi LU^2+1}} \approx \frac{1}{4U\sqrt{2\pi L}}$ when $LU^2 >> \frac{1}{8\pi}$.

The entropy of the equilibrium population *is* $LH(p)$ bits, where H is the entropy function $H(p) = -p\log_2 p - (1-p)\log_2(1-p)$. Very briefly, in this case the channel capacity is $Capacity = L(1-H(p)) \approx \frac{1}{16\pi(\ln 2)U^2} \propto \frac{1}{U^2}$ for p close to 1/2. Far more information can thus be maintained with large L and p close to ½ than with small L and p equal to 1.

3. Asexual reproduction

Very briefly, for asexual reproduction with the same selection rules and selection intensity, the analysis is similar, except that the variance of the number of agreements is $O(L\,U)$. Solving for p, we obtain $p - ½ = O((LU)^{-½})$, so that the amount of information is $O(U^{-1})$ whether for exact or diffuse encodings.

Discussion

This simple abstract analysis gives two striking and surprising results. First, in sexual reproduction, the manner in which information is encoded strongly affects the amount of information that can be maintained. In principle, vastly more information may be maintained at equilibrium if information is encoded diffusely in very long genomes. For such diffuse encoding, the amount of information that can be stored is proportional to U^{-2}, whereas if genetic information is encoded as an exact sequence of 0s and 1s, the amount that can be stored is proportional to U^{-1}, which is very much less for small U.

Second, in asexual reproduction the maintainable information is proportional to U^{-1}, even for diffuse encodings.

These results are obtained by a rather abstract argument and must be interpreted with care: they are upper limits on the amount of information that could conceivably be maintained in equilibrium, using the most favourable possible genetic

encodings and the most efficient possible selection rules. Nevertheless, where channel capacity exists it is likely to be used. The genomes of sexual eukaryotes are in general larger than those of the asexual prokaryotes: it is tempting to speculate that the "junk" DNA of sexual eukaryotes may encode some useful information in a very diffuse way, whereas the genomes of asexual prokaryotes are compact since there can be no advantage to a diffuse encoding.

Acknowledgements

We acknowledge helpful conversations with David McAllester and Quaid Morris.

References

1. Baum E.B., Boneh D., Garrett C. (1995) On Genetic Algorithms, COLT 95: Proceedings of the Eighth Annual Conference on Computational Learning Theory, ACM, New York. 230-239.
2. Cover T.M., Thomas J.A. (1991) Elements of Information Theory, Wiley-Interscience, New York.
3. Crow J.F., Kimura M. (1970) An Introduction to Population Genetics Theory, Harper and Row, New York.
4. Crow J.F., Kimura M. (1979) Efficiency of Truncation Selection. Proc. Natl Acad. Sci. USA. 76:396-9.
5. Kondrashov A.S. (1988) Deleterious Mutations and the Evolution of Sexual Reproduction. Nature. 336 (6198): 435-440.
6. Mackay D.J.C. (1999) Rate of Acquisition of Information of a Species subjected to Natural Selection, unpublished, available from http://www.mrao.cam.ac.um/~mackay
7. Muhlenbein H., Schlierkamp-Vosen D. (1993) Predictive Models for the Breeder Genetic Algorithm 1. Continuous parameter optimisation. Evolutionary Computation. 1: 25-50.

METHOD OF HORIZONTAL GENE TRANSFER DETERMINATION USING PHYLOGENETIC DATA

Lyubetsky V.A. [*], ***V'yugin V.V.***

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

Key words: *evolution, phylogenetic methods, trees consensus, horizontal gene transfer, mathematical models of evolution*

Resume

Motivation: An algorithm for comparative analysis of multiple trees reconstructed for representative protein families are discussed. This algorithm is based on the hypotheses of gene loss and horizontal gene transfers and uses stochastic methods and optimization. Some practical results are discussed. We describe a species tree comprising 40 prokaryotic organisms constructed by our algorithm on the basis of 132 individual groups of orthologous proteins (COGs) from GenBank of the National Center for Biotechnology Information (USA). We also describe a method for determination of horizontally transferred genes and its practical applications.
Results: An algorithm for horizontal gene transfer determination is developed. Several horizontally transferred genes were detected using this algorithm.
Availability: The software is available on request from the authors.

Introduction

The availability of numerous complete genome sequences from diverse taxa induces the development of new phylogenetic approaches, which incorporate information derived from comparative analysis of large gene sets. We consider two closely related approaches to reconstruction of phylogenetic relationships between species and to determination the horizontal gene transfer—the events occurring on molecular level-using these phylogenetic data.
An algorithm (V'yugin, Lyubetsky, 2001; V'yugin et al., 2002) for reconstructing phylogenetic species trees is described. This algorithm uses a set of possibly contradictory gene (protein) trees as initial data and produces a species tree as a census for these gene trees. The algorithm uses a model of gene duplications and losses explaining and measuring the dissimilarity between single gene tree and species tree (V'yugin, Lyubetsky, 2001). We also apply this model to determination of molecular events of horizontal gene transfer. Several computer experiments were realized using this algorithm. One of them is based on the data obtained from the National Center for Biotechnology Information, USA (Wolf et al., 2001). The data includes 132 maximum-likelihood trees constructed on the basis of 132 clusters of orthologous groups of proteins (COGs). The genomic sequences extracted belong to 40 living organisms from 13 groups: Archae (10), gamma-proteobacteria (7 organisms), gram-positive bacteria (8), alpha-proteobacteria (3), epsilon-proteobacteria (2), *Chlamydia* (2), spirochetes (2), beta-proteobacteria (1), cyanobacteria (1), *Deinococcus radiodurans* (1), *Aquifex aeolicus* (1), *Thermotoga maritima* (1), and *Mycobacterium tuberculosis* (1). The corresponding 132 maximum-likelihood protein trees were used as initial data in our method for constructing a census species tree. Species tree S is constructed as a tree closest to the 132 gene trees G_n. The measure of similarity is represented by a cost functional F. Definition of this functional is based on a model of evolution (V'yugin, Lyubetsky, 2001).

Methods and Algorithms

A stochastic algorithm of species tree reconstruction. The algorithm is based on a hypothesis of that the dissimilarity between gene trees constructed for different protein families results from a lineage-specific gene losses and duplications and horizontal gene transfer. We define a natural homomorphism (embedding) of gene tree into species tree and compare the gene and species trees by cost of this embedding. The value of functional F represents this cost (V'yugin, Lyubetsky, 2001). Following the principle of Occam's razor, we find a species tree minimizing this cost F of embedding (i.e. we try to find a species tree minimizing the total cost of molecular events of gene duplications and losses during the evolution of species). The search algorithm runs on a set of randomly generated (for example, 1000) initial species trees S_0. Any such tree S_0 is transformed using the method of nearest neighbor interchange to obtain a local minimum of the functional F. To specify this local minimum, we use an *a priory* probability distribution in the set of initial species trees. This probability distribution is defined as follows using 132 gene trees. Recall first that a distance between two leaves a and b is defined as a number of edges in the path between them. For any species a, an empirical probability distribution $p(b|a)$ that species b defines with a an elementary two-elements tree (i.e. they are located at a distance 2). We could define more detailed conditional distributions (for small trees of species located at a distance 3, 4, and so on), but this requires larger sets of initial data. The needed empirical probability distribution is defined using statistics of distribution of genes in 132 COG

[*] Corresponding author

trees. Let N_a be the number of COGs containing species a, and let $N_{a,b}$ be the number of COGs containing a and b located at a distance 2. We define $p(b|a) = N_{a,b}/N_a$. Then $1 - S_b\, p(b|a)$ is a probability that a forms a one-element elementary tree (i.e. no species locates at a distance 2 from a). The initial species a is defined using a uniform pseudorandom number generator, its neighbor b is defined using a generator of conditional probability distribution $p(. \mid a)$. We repeat this procedure to generate the next pair and so on. A random binary tree S_0 is generated on the basis of these elementary trees. This tree S_0 serves as the initial tree for the algorithm searching for optimal tree that gives a local minimum to the cost functional F. A variety of initial trees generate a variety of resulting trees produced by the search algorithm. As a final result, we output a consensus tree computed on the basis of a subset of these trees with sufficiently small values of the functional F. The numbers reflecting the reliability of corresponding clusters are assigned to the edges of this consensus tree. A resulting species tree explaining the evolutionary phylogenetic relationships of 40 living organisms (listed above) was constructed. This tree is close to a tree obtained in (Wolf et al., 2001) by comparative analysis of multiple trees reconstructed for representative protein families (approach (v) Wolf et al., 2001). Our species tree has very good suggestion for pairs of leaves and sufficiently good suggestion for the main 11 groups of organisms listed above. The difference between our tree and the best species tree from (Wolf et al., 2001) is only in the relative position of epsilon-proteobacteria group and the (Aae, Tma)-pair.

A method for detection of horizontally transferred genes. Horizontal gene transfer is a transfer of genes between organisms without reproduction. There are several hypotheses concerning the mechanisms of this transfer, for example, DNA can be transferred by infected bacteriophages or via mating mediated by plasmides (Lorencz, Wackernagel, 1994). In this section, we describe a method for detecting the genes suspected of being horizontally transferred. This method is based on a hypothesis implying that an event of horizontal gene transfer involves essential dissimilarity between the gene and species trees. We use two methods for estimating this dissimilarity. First is based on comparison of neighborhoods of a gene in the gene tree and its image in the species tree. If two genes are located at a small distance in the gene tree but their images are dispersed in the species tree, this indicates a possible horizontal transfer of one of them. We measure dispersion of a gene neighborhood under the homomorphism of gene tree into species tree. Let v1, v_2, ..., v_n be all the genes located in the neighborhood of gene v of a radius r, and let s_1, s_2, ..., s_n and s be the corresponding species (their images in species tree). The distances $r(v,v_i)$ in the gene tree and distances $r(s,s_i)$ in the species tree are calculated, where, i =1, …, n. We also calculate the average values as

$$r(v) = (1/n)\, S_i\, r(v,v_i) \text{ and } r(s) = (1/n)\, S_i\, r(s,s_i).$$

The ratio $p = r(s)/r(v)$ reflects the degree of average dissipation of the gene v vicinity in the species tree. Large values of this ratio can be interpreted as reflecting pathology in the location of gene v in the species tree. The computer program outputs the list of the genes suspected of being horizontally transferred. Each gene in this list is supplied with certain confidence information. In our analysis, we also take into account the diversity of COG (gene) trees. A high cost of COGs embedding into species tree reduced the confidence level of the results concerning genes from this COG.

The second approach uses a hypothesis that the temporary deletion of a transferred gene from the gene tree and updating of its image in the species tree after this deletion implies essential decrease in the value of cost functional F (for example, the cost can be decreased by many sigma from a mean). To apply this method, a normalization of the gene trees is needed. The edges of maximum likelihood gene trees are supplied with the numbers (lengths) reflecting the time of evolution of corresponding genes. The longest lengths have the biggest impact on the total value of functional F. This, wrong locations of these longest edges (which can result from inaccuracy of the maximum likelihood method) bring about the biggest error to the value of the functional. To eliminate this effect, we normalize excessively long leaf edges in the gene tree. For any COG tree, the cost F of embedding of the corresponding tree in the species tree is calculated. We remove temporarily each gene g from gene tree G to obtain a reduced gene tree G_g and compute the cost F_g of embedding the gene tree G_g in the species tree. The relative change in the cost of embedding is calculated as $dF_g = (F_g - F)/F$. We sort all the genes from the COG by absolute values of dF_g. We suppose that the order numbers of genes suspected of being horizontally transferred has a high correlation with large absolute values of dF_g. More correctly, the mean and variance of dF_g were used to compute the confidence information for each gene from any COG. Additional confidence information was computed as follows. For each gene g, a value

$$dF_{cp}^{\,r}(g) = (1/n(r))\, S_{g:r(s,g)<r,\ s \neq g}\, dF_s$$

was calculated, where $n(r)$ is the number of leaves in the neighborhood of g of a radius r. In our experiments, we used r = 1, 2, 3, 4, and 5. A large value of

$$k_g^{\,r} = (dF(g))/(\sum dF_{cp}^{\,r}(g))$$

suggests that gene g in the species tree displays a pathological location.

This approach can be applied more efficiently to a case of horizontal transfer of groups of genes. Combined lists of genes suspected of being horizontally transferred were formed on the basis of both methods. Each gene in the list is supplied with confidence information. This information can serve as a tool for an expert analyzing the molecular events in the process of evolution.

Implementation and Results

In this section, we present some results of detecting genes suspected of being horizontally transferred in the descending order of their reliability levels.

1) *yicF* is a gene in the COG0272 extracted from the genome of *E. coli* (gamma-proteobacteria group). The gene was selected as (a) a temporary deletion of it from the gene tree gives a large deviation of the value dF_g from the mean value of this variable (18 sigma) and (b) the nearest neighbors of this gene in the gene tree embed into group of gram-positive bacteria located at a long distance from gamma-proteobacteria in the species tree (the ratio $p = r(s)/r(v)$ defined above is equal to 8). We suppose that the group of gram-positive bacteria is the source of this horizontal transfer.
2) *aq946* is a gene in the COG0571 extracted from the genome of *Aquifex aeolicus;* it also displays a large deviation of the value dF_g from the mean value of this variable (15 sigma), and the nearest neighbors of this gene in the gene tree embed into group of alpha-proteobacteria. This group is supposed to be the source of this horizontal transfer.
3) *VNG1097G* is a gene in the COG0215 extracted from genome of *Halobacterium sp.* NRC-1 (the family Archae). A temporary deletion of this gene from the gene tree gives a large deviation of the value dF_g from the mean value (10 sigma), and the nearest neighbors of this gene in gene tree embed near *Deinococcus*, which is located in the species tree at a long distance from Archae.
4) *RP687* is a gene in the COG0525 extracted from the genome of *Rickettsia prowazekii* (the group of alpha-proteobacteria). The nearest neighbors of this gene in gene tree embed near Archae, which is located in the species tree at a long distance from alpha-proteobacteria group. A deletion of this gene also changes essentially the value of the functional.
5) *VNG2507G* is a gene in COG0167 extracted from the genome of *Halobacterium sp.* NRC-1 (the group of Archae). The nearest neighbors of this gene in gene tree embed near the group of *Mycobacterium tuberculosis, Synechocystis,* and *Deinococcus radiodurans,* which is located in the species tree at a long distance from Archaea. A deletion of this gene also changes essentially the value of the functional.

References

1. Lorencz M.G., Wackernagel W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. Microbial Reviews. 58:563-602.
2. V'yugin V.V., Gelfand M.S., Lyubetsky V.A. (2002). Trees reconciliation: species trees reconstruction by phylogenetic gene trees. Mol. Biol. (accepted for publication).
3. V'yugin V.V., Lyubetsky V.A. (2001). On algorithm of horizontal gene transfer searching based on phylogenetic protein trees. Informatsionnye Protsessy. 1(2):167-177.
4. Wolf Y., Rogozin I., Grishin N., Tatusov R., Koonin E. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. BMC Evol. Biol. 1:8.

MINIMAL TREES IN PHYLOGENETIC SPACES

Ivanov A.O., Tuzhilin A.A.

Faculty of Mechanics and Mathematics, Moscow State University, Moscow, Russia, e-mail: aoiva@mech.math.msu.su; tuz@mech.math.msu.su

Key words: *phylogenetic trees, Steiner minimal trees, editorial distance, Levenstein metric space*

Resume

Motivation: To apply the ideas of Minimal Networks Theory to phylogenetic trees constructing problem.
Results: We give an algorithm constructing a minimal tree Γ of a given topology G spanning a finite subset $N=\{\beta_1,...,\beta_n\}$ of a phylogenetic pseudo-metric space. The velocity of the algorithm has the order of $2^n|\beta_1|\cdots|\beta_n|$, where $|\beta|$ stands for the pseudo-length of a word β. As a corollary, we obtained algorithms constructing a Simpson–Torricelli point for the set N and, in particular, solving Steiner problem for three points boundary set in a phylogenetic space. Notice that the work is a direct generalization of classical results obtained by D.Sankoff (1975), where phylogenetic spaces with Levenstein metric were considered. The present work was done in collaboration with Professor D. Cieslik, Germany.

Introduction and Necessary Definitions

Phylogenetic spaces are metric spaces whose points are arbitrary words generated by letters from some finite alphabet, and metric measuring "sameness" of the words is generated by a distance function given on the letters. Phylogenetic spaces theory has many applications. For example, in biological Evolution Theory, see Dress et al. (1986), the words are constructed from the letters corresponding either to aminoacids generating proteins, or to the nucleotides forming DNA and RNA. Comparing such words obtained from investigations of specific living organisms one can construct evolutionary (phylogenetic) trees showing a "relation degree" of the species considered: closeness of the words in the tree corresponds to the closeness of the species. The latter remark explains the importance of trees having the least possible length in phylogenetic spaces for evolutionary relation investigation: appearance of such trees demands the least number of evolution's steps H.-J. Bandelt, et al. (1995). This approach to Evolution theory was suggested first by W.Fitch (1971). Another important application of phylogenetic spaces theory is investigation of languages' evolution.

Let $A=\{a_1,...,a_m\}$ be a finite set. The elements a_i of A are called *letters*, and the set A itself is called an *alphabet*. Let A^* be the set of all finite sequences (including the empty one) composed from the letters a_i. Elements of A^* are called *words*. The word without letters is said to be *empty* and is denoted by 0. Notice that sometimes it is convenient to consider 0 as the *empty letter* also. The number of (non-empty) letters in a word α is called the *length of the word* and is denoted by $|\alpha|$. In particular, $|0|=0$. We write word α of the length k as $\alpha=\alpha^1\cdots\alpha^k$, where $\alpha^i\in A$.

Let us put $A_0=A\cup\{0\}$, and let ρ be some pseudo-metric defined on A_0 (recall that a non-negative function $\rho:X\times X\to R$ is called a *pseudo-metric on X* if $\rho(x,y)=\rho(y,x)$ and $\rho(x,z)\le\rho(x,y)+\rho(y,z)$ for any x, y, and z from X; we do not demand the non-degeneracy property: the distance between distinct points can be equal to zero). Let us extend ρ onto the set A^*. To do that let us define so-called "editorial operations" named deletions, insertions, and substitutions. Let $\alpha=\alpha^1\cdots\alpha^k$ be some word from A^*. If p is a positive integer, and $x\in A$, then the operation taking the word α to

- the word $\beta=\alpha^1\cdots\alpha^{p-1}\alpha^{p+1}\cdots\alpha^k$ is called the *deletion* and is denoted by del_p, $1\le p\le k$;
- the word $\beta=\alpha^1\cdots\alpha^{p-1}x\,\alpha^p\alpha^{p+1}\cdots\alpha^k$ is called the *insertion* and is denoted by $ins_p[x]$, $1\le p\le k+1$ (note that the insertion can be done as before the word (p=1), as after the word (p=k+1));
- the word $\beta=\alpha^1\cdots\alpha^{p-1}x\alpha^{p+1}\cdots\alpha^k$ is called the *substitution* and is denoted by $sub_p[x]$, $1\le p\le k$, $x\ne\alpha^p$.

Let us assign a weight to each of editorial operations as follows:

- $\rho(0,\alpha^p)$ for del_p;
- $\rho(0,x)$ for $ins_p[x]$;
- $\rho(\alpha^p,x)$ for $sub_p[x]$.

If we are given with a sequence of the operations, then the total weight of them is called the *weight of the sequence*. Let α and β be some words from A^*. The greatest lower bound of the weights of all possible operations sequences taking α to β is called the *pseudo-distance between α and β* and is denoted by $\rho(\alpha,\beta)$. It is easy to see that this greatest lower bound is attained at some (finite) operations sequence (this sequence is said to be *ρ-realizing*), and that the pseudo-distance between the one-letter-words (including the empty one) coincides with the predefined pseudo-distance between them as

between elements of A_0. Moreover, it is easy to verify that ρ is a pseudo-metric on A^*. The pseudo-metric space (A^*,ρ) is called the *phylogenetic space generated by* (A_0,ρ).

Remark. If the function ρ is a metric on A_0, then the corresponding pseudo-metric on A^* is a metric also. For example, if we take $\rho(a,b)=1$ for any different elements of A_0, then the corresponding pseudo-metric on A^* is a metric and defines well-known Levenstein distance.

Let V be an arbitrary finite set. Recall that a *graph G on a set V* is a pair (V,E), where E is a finite family of pairs of elements from the set V. If G is a graph on V, and $W\subset V$, then we say that G *spans W*.

We will consider only *simple* graphs, that is, we assume the family E does not contain pairs of the same element (loops) and the same pairs of elements (multiple edges). Elements from V are called *vertices of G*, and elements from E are called *edges of G*. For a given graph G, the set of its vertices is usually denoted by V(G), and the set of its edges by E(G). For convenience, we often denote an edge of the form $e=\{x,y\}\in E(G)$ simply by xy.

Let X be a set, ρ be some pseudo-metric on X, and N be an arbitrary finite subset of X. Let G be a graph on N. The *length of G* is the number

$\rho(G)=\sum_{xy\in E(G)}\rho(x,y)$.

Further, we put

$MST^X(N)=\min_G \{ \rho(G) \mid G \text{ is a tree, } V(G)=N \}$,

$SMT^X(N)=\inf_G \{ \rho(G) \mid G \text{ is a tree, } N\subset V(G)\subset X \}$,

$$SiMT^X(N)=\inf_{y\in X}\sum_{x\in N}\rho(y,x).$$

Let G be a graph, V be the set of its vertices, and $\Gamma:V\to X$ be an arbitrary mapping. If the graph G is connected, then Γ is called a *network* in the space X. If xy is an edge in G, then the mapping $\gamma:xy\to\Gamma(x)\Gamma(y)$ is called an *edge of the network* Γ. The *pseudo-length* $\rho(\gamma)$ *of the edge* $\gamma:xy\to\Gamma(x)\Gamma(y)$ is defined as $\rho(\Gamma(x),\Gamma(y))$, and the *pseudo-length* $\rho(\Gamma)$ *of the network* Γ is the sum of the pseudo-lengths of all its edges.

The networks we are interested in are obtained as solutions to the following boundary value problem. Let us fix a subset ∂G of the vertex set V of the graph G and call it by a *boundary of* G. Let us fix an arbitrary mapping $\varphi:\partial G\to X$ and call it *boundary*. By $[G,\varphi]$ we denote the set of all the networks $\Gamma:V\to X$ such that $\Gamma|_{\partial G}=\varphi$. Vertices belonging to the boundary are said to be *boundary*, and the remaining ones are called *interior* or *mobile*. An edge incident to a boundary vertex is called *boundary* also. All the remaining edges are said to be *interior* or *mobile*.

Let G be a tree. We put

$PMT_G(\varphi)=\inf_{\Gamma\in[G,\varphi]}\rho(\Gamma)$.

A tree G on N is said to be a *minimal spanning tree* if $\rho(G)=MST^X(N)$. A tree G on a finite subset $N\subset X$ containing N is called a *Steiner minimal tree spanning N* if $\rho(G)=SMT^X(N)$. A point $y\in X$ such that $\sum_{x\in N}\rho(y,x)=SiMT^X(N)$ is called a *Torricelli point*. At last, a network $\Gamma\in[G,\varphi]$ such that G is a tree and $\rho(\Gamma)=PMT^X{}_G(\varphi)$ is said to be a *parametric minimal tree of the type* $[G,\varphi]$.

Notice that a minimal spanning tree exists for an arbitrary set N, but Steiner minimal trees, parametric minimal trees, Torricelli points may not exist, see example in A.O.Ivanov, A.A.Tuzhilin (2001).

Main Results

Let G be an arbitrary tree with a boundary $G=\{b_1,...,b_n\}$, and let $\{s_1,...,s_k\}$ be remaining (mobile) vertices of the graph G. Consider an arbitrary *boundary* mapping $\varphi:\partial G\to A^*$ and put $\beta_i=\varphi(b_i)$. In what follows it is convenient to denote $PMT^{A^*}{}_G(\varphi)$ simply by $PMT^{A^*}{}_G(\beta_1,...,\beta_n)$.

Further, let $\mathbf{B}=\{0,1\}$. For $\varepsilon\in\mathbf{B}$ we put $\varepsilon'=1-\varepsilon$. If $x\in A_0$, and $\varepsilon\in\mathbf{B}$, then let $x^\varepsilon=x$ for $\varepsilon=1$, and $x^\varepsilon=0$ for $\varepsilon=0$.

Recall that the set $\mathbf{B}^n$ is endowed with the following natural partial order: for $\varepsilon=(\varepsilon_1,\cdots,\varepsilon_n)$ and $\delta=(\delta_1,\cdots,\delta_n)$ from $\mathbf{B}^n$ we put $\varepsilon\le\delta$ if and only if $\varepsilon_i\le\delta_i$ for all i=1,...,n, and $\varepsilon<\delta$ if and only if $\varepsilon\le\delta$ and $\varepsilon\ne\delta$. For $\chi\in\mathbf{B}^n$ we put $\mathbf{B}^n{}_\chi=\{\varepsilon\in\mathbf{B}^n \mid \varepsilon<\delta\}$. If $\chi=(1,...,1)$, then by $\mathbf{B}^n{}_1$ we denote $\mathbf{B}^n{}_\chi$ for brevity. Notice that $\mathbf{B}^n{}_\chi=\varnothing$ if and only if $\chi=(0,...,0)$, and $\mathbf{B}^n{}_\chi\subset\mathbf{B}^n{}_1$ for any χ.

Let $x=(x_1,...,x_n)$ be a sequence of letters from A_0. We define the *indicator* $\chi(x)=(\chi_1,...,\chi_n)\in\mathbf{B}^n$ as follows: $\chi_i=0$ if and only if $x_i=0$.

Theorem 1. Let $x_1,...,x_n\in A_0$ be arbitrary letters not equal to 0 simultaneously, and let $\beta_1,...,\beta_n\in A^*$ be words such that if $x_i=0$, then $\beta_i=0$. Put $\chi=\chi(x_1,...,x_n)$. Then for any tree G with a boundary $\{b_1,...,b_n\}$ and for the boundary mapping $\varphi:b_i\to\beta_i x_i$ the equality holds

$$PMT^{A^*}{}_G(\beta_1x_1,...,\beta_nx_n)=\min_{\varepsilon\in\mathbf{B}^n{}_\chi}[PMT^{A^*}{}_G(\beta_1x_1^{\varepsilon_1},...,\beta_nx_n^{\varepsilon_n})+PMT^{A_0}{}_G(x_1^{\varepsilon'_1},...,x_n^{\varepsilon'_n})],$$

where $\varepsilon=(\varepsilon_1,...,\varepsilon_n)$.

Theorem 1 gives an opportunity to calculate the pseudo-length of a minimal parametric tree of a given type with a fixed boundary. The next Theorem permits to find mobile vertices of this tree.

Theorem 2. Under the assumptions of Theorem 1, let the pseudo-length of a parametric minimal tree of the type G have the form

$$PMT^{A^*}_G(\beta_1 x_1,...,\beta_n x_n)=PMT^{A^*}_G(\beta_1 x_1^{\varepsilon_1},...,\beta_n x_n^{\varepsilon_n})+PMT^{A_0}_G(x_1^{\varepsilon'_1},...,x_n^{\varepsilon'_n})$$

for some $\varepsilon=(\varepsilon_1,...,\varepsilon_n)\in \mathbf{B}^n_1$. Let σ_i and t_i, i = 1,...,k, be the corresponding to $s_i\in V(G)$ mobile vertices of parametric minimal trees of the type G with the boundaries $(\beta_1 x_1^{\varepsilon_1},...,\beta_n x_n^{\varepsilon_n})$ and $(x_1^{\varepsilon'_1},...,x_n^{\varepsilon'_n})$, respectively. Then the parametric tree Γ of the type G such that $\Gamma(b_i)=\beta_i x_i$ and $\Gamma(s_i)=\sigma_i t_i$ is minimal.

These two theorems imply directly result concerning Steiner minimal trees and Simpson-Torricelli points mentioned in the Resume. The details can be found in D.Cieslik, A.O.Ivanov, A.A.Tuzhilin (2002).

References

1. Ivanov O., Tuzhilin A.A. (2001) Calculus on the space of Steiner minimal trees in Riemannian manifolds. Matem. Sb. 192, 6:31-50.
2. Cieslik D., Ivanov A.O., Tuzhilin A.A. (2002) Minimal Trees in Phylogenetic Spaces. Vestnik MGU. N 3.
3. Bandelt H.-J. et al. (1995) Mitochondrial Portraits of Human Populations Using Median Networks. Genetics. 141:743–753, 1995.
4. Dress et al. (1986) Reconstructing Phylogenetic Trees using Variants of the "Four-Point-Condition". Studien zur Klassifikation. 17:299–305.
5. Fitch W. (1971) Toward defining the course of evolution: minimum change for specific tree topology. Systematic Zoology. 20:406–416.
6. Sankoff D. (1975) Minimal mutation trees of sequences. SIAM J. of Appl. Math. 28, 25–42.

COMPARATIVE ANALYSIS OF CODING SEQUENCES OF *APETALA1* HOMOLOGUES

Omelyanchuk N.A. *[1], Gusev V.D. [2], Nemytikova L.A. [2], Aksenovich A.V. [1]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: nadya@bionet.nsc.ru
[2] Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia

Key words: *homologues, flower development*

Resume

Motivation: So far, numerous homologues of an *Arabidopsis* MADS box gene *APETALA1* (AP1) have been sequenced. Detection of conservative regions of *AP1* gene homologues at the DNA level is important for the understanding of regulatory role of these genes and their practical application in genetic engineering.
Results: Comparative analysis of the coding regions of *AP1* gene and its homologues allowed us to detect the regions conservative for certain gene groups. These regions are enriched with binding sites for regulatory factors involved in chromatin modulation. Calculation of the mRNA secondary structure with a least energy has demonstrated that the regions in question correspond to the complementary sequence fragments forming the stem elements in mRNA secondary structures. Thus, the conservation of coding sequences of *AP1* gene and its homologues may result not only from the conservation of the corresponding protein sequences, but also from the conservation of the regions forming stem elements in the RNA secondary structure and conservation of the DNA regions probably involved in chromatin modulation.

Introduction

In *Arabidopsis,* the gene *APETALA1* (*AP1*) promotes flower meristem identity and provides the development of sepals and petals. *AP1* belongs to the family of MADS box floral homeotic genes (Mandel et al., 1992). These genes are structurally similar, and their products show high level of homology in some regions (Theissen et al., 1996). In plants, the MADS box proteins consist of four separate domains: M (MADS domain), I (intervening region, a linker between the MADS domain and the K box), K (K box), and C (C-terminal) regions (Purugganan et al., 1995). Among MADS box protein domains, the MADS domain is highly conserved and involved in DNA binding (Riechmann and Meyerowitz, 1997; Theissen et al., 1996). The goal of this work was to detect and study the conservative regions of the gene *AP1* coding sequence and to search for potential explanations of the presence of these conservative regions in the gene *AP1*.

Methods and Algorithms

The sample studied was extracted from GenBank and EMBL and supplemented with additional sequences through searching for homologues to particular regions of gene *AP1* using BLAST (http://www.ncbi.nlm.nih.gov/BLAST/). Gene fragments with a length $l = 15$ identical to certain gene *AP1* regions were found using an algorithm for detecting interconnections in sets of functionally and/or evolutionary related biological sequences (Gusev et al., 2001), realized as a software system LZcomposer http://wwwmgs.bionet.nsc.ru/programs/lzcomposer (Gusev et al., 2001). Secondary structure of mRNA was calculated using the program GArna http://wwwmgs.bionet.nsc.ru/mgs/programs/2dstructrna/. Nucleosomal potential was determined using the program Recon http://wwwmgs.bionet.nsc.ru/mgs/programs/recon/. Potential GATA transcription factor binding sites were detected by the program MatInspector V2.2 http://transfac.gbf.de/cgi-bin/matSearch/matsearch.pl.

Results and Discussion

MADS region contains the highest number of conservative fragments. In 320 partial and complete coding sequences, the fragments identical to gene *AP1* with a length of 15 nt and more were found in all the regions of these gene sequences; however, they were most abundant in the region coding for MADS box (MADS region). Conservative fragments found in this region fell into four groups and divided the MADS region into four individual separate subregions with reference to their localization, namely M1 (1–18), M2 (19–57), M3 (58–135), and M4 (136–171). Indicated in the parentheses are positions of the first and last nucleotides of the corresponding regions. The extreme regions M1 and M4 displayed the least numbers of the fragments found. The M4-contained fragments identical to gene AP were detected only in AP1 homologues of several Cruciferae species. As for the M1 subregion, 18-nt long identical regions, in addition to Cruciferae, were found in two species belonging to other families.
The central subregions M2 and M3 are the most conservative in MADS region. The fragments identical to M2 sequence fragments were found in 56 genes belonging to plants of 19 genera of various families (Fig. 1). Among remote homologues, identical fragments were found in genes *GGM9* and *GpMADS3* of representatives of the genus *Gnetum,* belonging to gnetophytes. Fragments identical to parts of subregion M3 were found in 78 genes of plants from 36 genera.

a

M1: from 1 to 18 nucleotide	M2: from 19 to 57 nucleotide	M3: from 58 to 135 nucleotide	M4 from 136 to 171 nucleotide

b

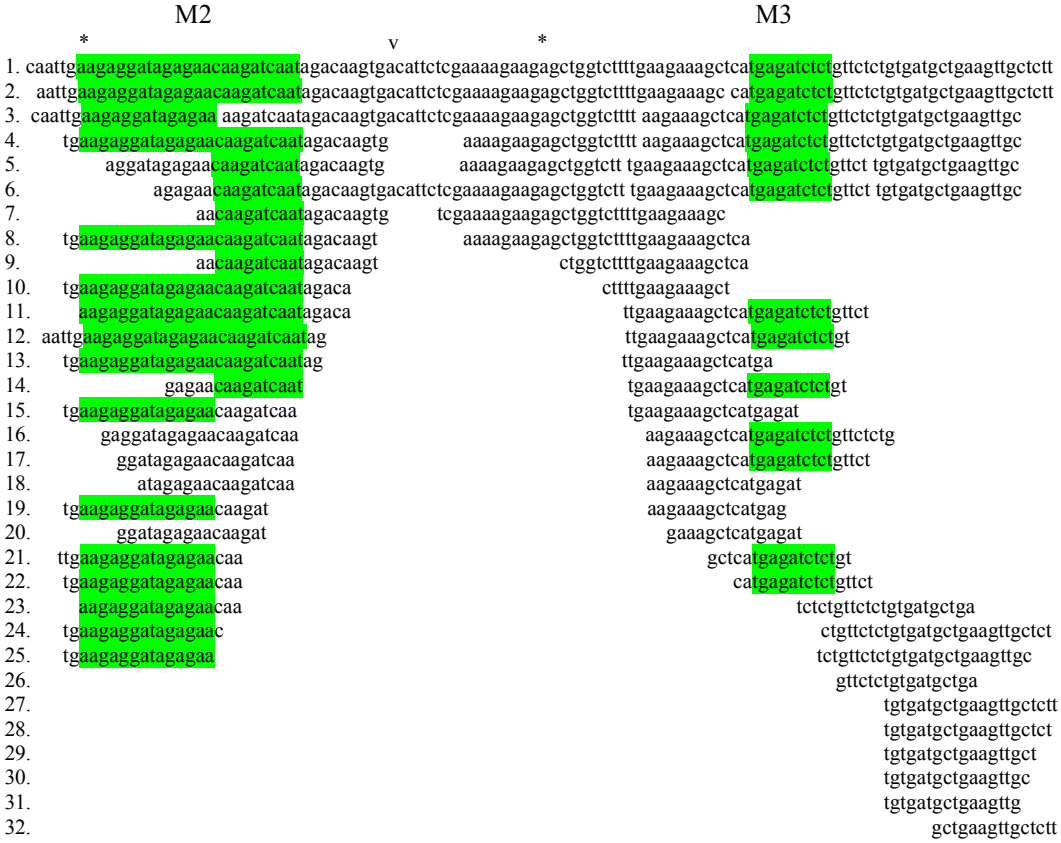


Fig. 1. Characterization of gene *AP1* MADS region with conservative fragments. (**a**) Partition of MADS region into four subregions according to localizations of conservative fragments and (**b**) Fragments of homologous genes identical to *AP1* M2 and M3 regions: **M2**: 1. *AP1* (At, Al), *C-2* (Sin) 2. *CAL* (At) 3. *SAAP1* (Sin) 4. *BOI1AP1, BOBCAL* (Br) 5. *BRPCAL* (Br) 6. *CAL* (Al) 7. *BOAP1, Boi2AP1* (Br) 8. *AGL3* (At) 9. *SLM4* (Sil) 10. *SQUA* (An) 11. *MADS AF002254* (Or) 12. *AGL9* (At), *SAMADSD* (Sin) 13. *AGL8* (At) 14. *NAP1-1, NsMADS1* (Ni), *SLM5* (Sil), *GGM9, GpMADS3* (Gn) 15. *BOCAL, BOICAL* (Br), *MADS3*(Vi), *EAP*1 (Eu) 16. *HAM75* (Hel) 17. *MADS15, MADS AB003325* (Or) 18. *PMADS4* (Pe) *NSMADS3* (Ni), *MdMADS5* (Mal), *DEFH72, DEFH200* (An) 19. *AGL6* (At), *MAGL4* (Po) *EAP2S, AP2L* (Eu), *MDMADS1, MDMADS8*, *MDMADS9* (Mal), *GSQUA* (Ge) 20. *SBMADS2* (Sor), *MADS2* (Lo), *M8* (Ho) 21. *CAGL2* (Cu) 22. *AGL2* (At), *MADS2* (Vi), *PrMADS1* (Pin), *EGM3* (Eu) 23. *AGL4* (At), *FBP9* (Pe) 24. *AGL14* (At) and 25. *MDMADS3, MDMADS7* (Mal); **M3**: 1. *AP1* (At), *AP1 male* (Po) 2. *AP1* (Al) 3. *BOAP1* (Br) 4. *BOIAP1* (Br) 5. *AP1* (Sin) 6. *C-2* (Sin), *BOI2AP1* (Br) 7. *BOCAL, BOBCAL, BOICAL, BRPCAL* (Br) 8. *CAL* (Al) 9. *CAL* (At) 10. *AGL14* (At), *CAGL2* (Cu), *TDR5* (Ly), *DEFH72, DEFH200* (Pe) 11. *GSQUA1* (Ge) 12. *MADS6* (Ca) 13. *AP1 female* (Po), *PFG* (Pe) 14. *SGM1MADS, POTM1-1, POTM1-2* (Sol) 15. *MDMADS2* (Mal) 16. *AGL8* (At), *SAMADSB* (Sin) 17. *PEAM4, MADS 4* (Pis) 18. *SLM5* (Sil), *MADS4* (Be) 19. *AGL15* (Br), *ASAP3/TM6* (Asteraceae). 20. *FBP26* (Pe) 21. *MDMADS5* (Mal) 22. *MADS3* (Be) 23. *EGM3* (Eu), *STAG1* (Fr), *PRMADS1* (Pin), *AG* (At) 24. *AGL2* (At) 25. *FBP23* (Pe) 26. *MDMADS1, MDMADS4, MDMADS8*, *MDMADS9* (Mal), *MADS1* (Hem), *MADS1* (Co), *CerMADS2* (Ce) 27. *MADS1* (Ca), *CAGL2*(Cu), *FBP2*, *FBP9* (Pe), *CMB1* (Di), *SLM4* (Sil), *MDMADS14* (Mal), *D1* (Ro), *NtMADS4* (Ni) 28. *PMADS4* (Pe), *CAG3*, *CAG1, CUM1, CUM10, M1, CMB1* (Cu), *TM29* (Ly) 29. *NAP1-2, NSMADS2* (Ni), *PEAM4* (Pis), *HAM75* (He), *GSQUA1* (Ge), *MADSMC* (Ly), *FBP22* (Pe) 30. *AGL14, AGL15, AGL19* (At), *SHP1* (Br), *FBP13*, *FBP6, PAGL1 PMADS3*, (Pe), *NSMADS2* (Ni), *FAR* (An), *TDR8* (Ly), *SLM1* (Sil), *MDMADS10* (Mal), *MpMADS11* (Mag) 31. *MADS6* (Ca), *PEGMADS, PFG, FBP26* (Pe), *CMB* (Cu) *PEAM4* (Pis) and 32. *FBP5* (Pe). Parenthesis contain abbreviated designations of genera and species: *Antirrhinum* (An), *Arabidopsis thaliana* (At), *A. lyrata* (Al), *Betula* (Be), *Brassica* (Br), *Capsicum* (Ca), *Ceratopteris* (Ce), *Corylus* (Co), *Cucumis* (Cu), *Dianthus* (Di), *Eucalyptus* (Eu), *Fragaria* (Fr), *Gerbera* (Ge), *Gnetum* (Gn), *Helianthus* (Hel), *Hemerocallis* (Hem), *Hordeum* (Ho), *Lolium* (Lo), *Lycopersicon* (Ly), *Magnolia* (Mag), *Malus* (Mal), *Nicotiana* (Ni), *Oryza* (Or), *Petunia* (Pe), *Pinus* (Pin), *Pisum* (Pis), *Populus* (Po), *Rosa* (Ro), *Silene* (Sil), *Sinapis* (Sin), *Solanum* (Sol), *Sorghum* (Sor), and *Vitis* (Vi).

Arrow indicates the beginning of M3 region; asterisks, boundaries of the region encoding a putative nuclear localization signal. Potential GATA4, GATA3, and GATA2 binding sites are filled.

The range of remote homologues containing the fragments identical to parts of M3 is wider compared with M2 and includes, for example, the fern gene *CerMADS2* along with homologues of genes *AGAMOUS* and *APETALA3*. The sequence coding for a putative nuclear localization signal, which is highly conserved in MADS box at the protein level, occupies almost all M2 subregion, starting from nucleotide 25 till the end and further reaches nucleotide 75 in M3 (Mandel et al., 1992). As is evident from Figure 1, the conservative regions of the sequence encoding a putative nuclear localization signal at the DNA level cluster mainly in its first half.

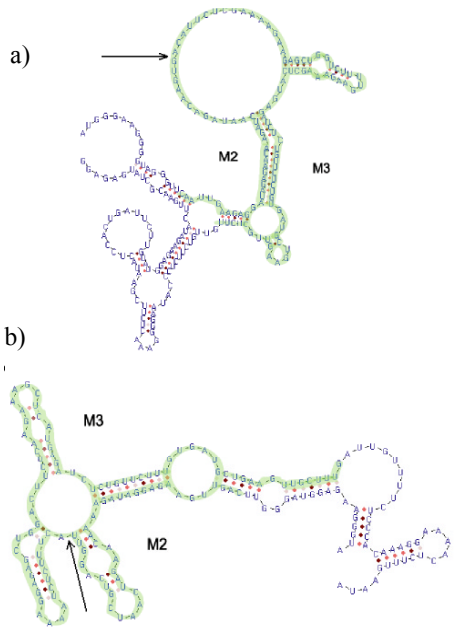


Fig. 2. Fragments of predicted mRNA secondary structures of (**a**) *Arabidopsis thaliana* gene *AP1* and (**b**) *Pisum sativum* gene *PEAM4*. The regions corresponding to M2 and M3 are colored. Arrow indicates the beginning of M3.

Conservative fragments of M2 and M3 regions correspond to particular elements of gene AP1 mRNA secondary structure. A least energy secondary structure of gene *AP1* mRNA was calculated. Shown in Figure 2a is a fragment of this model starting from the first nucleotide of MADS box and covering a part of the L region. As is seen from Figure 2, complementary interactions between the regions M2 and M3 form one of the stem elements in the mRNA secondary structure. Sequences homologous to the subregions M2 and M3 contained in other relevant genes may also be involved in formation of stem elements in mRNA secondary structure. This is illustrated with the fragment of predicted pea gene *PEAM4* mRNA secondary structure corresponding to MADS region (Fig. 2b). It is evident that the sequences of M2 and M3 also interact in the secondary structure of *PEAM4* forming a stem element.

Localization of the protein sites modulating chromatin structure in M2 and M3 regions. It is known that GATA transcription factors are capable of modulating the structure of chromatin (Cirillo et al., 2002; Muro-Pastor et al., 1999). Analysis of *AP1* coding sequence with the program Recon demonstrated a high nucleosome potential of MADS region. The potential GATA4, GATA3, and GATA2 binding sites are shown filled in Figure 1. These sites were detected in 45 genes containing fragments identical to parts of gene *AP1* in their M2 region and 17 genes containing those fragments in M3 region.

Conclusion

Thus, the conservation of coding sequences of *AP1* gene and its homologues may result not only from the conservation of the corresponding protein sequences, but also from the conservation of the regions forming stem elements in the RNA secondary structure and conservation of the DNA regions probably involved in chromatin modulation.

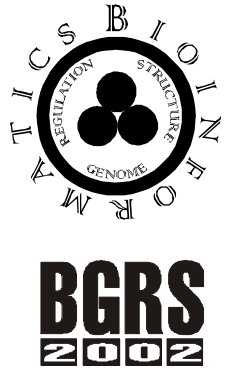
Acknowledgements

This work was supported in part by the Russian Foundation for Basic Research (grants № 00-04-49255, 01-07-90376, 00-07-90337, 02-07-90355, 00-04-49229, and 00-06-80420); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); and Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Cirillo L.A., Lin F.R., Cuesta I., Friedman D., Jarnik M., Zaret K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol. Cell. 9**,** 279-289.
2. Gusev V.D., Nemytikova L.A., Chuzanova N.A. (2001). A rapid method for detecting interconnection between functionally and/or evolutionary close biological sequences. Mol. Biol. (Mosk.). 35, 1015-1022.
3. Mandel M.A., Gustafson-Brown C., Savidge B., Yanofsky M.F. (1992). Molecular characterization of the *Arabidopsis* floral homeotic gene *APETALA1*. Nature. 360**,** 273-277.
4. Muro-Pastor M.I., Gonzalez R., Strauss J., Narendja F., Scazzocchio C. (1999). The GATA factor AreA is essential for chromatin remodelling in a eukaryotic bidirectional promoter. EMBO J. 18, 1584-1597.
5. Purugganan M. D. (1997) The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. J. Mol. Evol. 45, 392-396.
6. Riechmann J.L., Meyerowitz E.M. (1997) MADS domain proteins in plant development. Biol. Chem. 378**,** 1079-1101.
7. Theissen G., Kim J.T., Saedler H. (1996). Classification and phylogeny of the MADS-box multigene family suggest defined roles of MADS-box gene

subfamilies in the morphological evolution of eukaryotes. J. Mol. Evol. 43, 484-516.



SYSTEM COMPUTATIONAL BIOLOGY: ANALYSIS AND MODELING OF GENE NETWORKS AND METABOLIC PATHWAYS

GENENET SYSTEM: ITS STATUS IN 2002

**** Ananko E.A., Podkolodny N.L., Ignatieva E.V., Podkolodnaya O.A., Stepanenko I.L., Kolchanov N.A.***

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

Key words: *gene networks, signal transduction pathways, metabolic pathways, regulation of gene expression, visualization*

Resume

Motivation: A multitude of processes controlled by gene networks proceed simultaneously in cells, tissues, organs, and entire bodies. Recently developed technologies for studying gene expression provide rapid accumulation of information on a variety of molecular processes. As a rule, such information is only miscellaneous pieces of mosaic that fail to represent an integral pattern of the system function.
Results: We have developed the system GeneNet, allowing manifold data on gene networks to be accumulated in the context of a unified approach and to be visualized as interactive graphs. The system developed allows a wide diversity of gene networks to be described. The GeneNet now contains descriptions of 25 gene networks regulating such processes as lipid metabolism, functions of the immune and endocrine systems, responses to a number of external factors, etc. The information compiled in GeneNet was used to create dynamic models simulating the function of three gene networks.
Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenetworks.shtml

Introduction

Numerous molecular genetic, biochemical, and physiological processes run simultaneously in cells, tissues, organs, and entire organisms under the control of gene networks. A rapid development of modern technologies brings about a tremendous flow of diverse information on these processes. Various databases are created to arrange this information and make it available for computer analysis. As a rule, these databases compile information on particular sections of molecular biology. However, this is insufficient to reconstruct an integrated pattern of the gene network function.
Several essential structural and functional components are present in any gene network, namely, (1) concertedly expressed genes, forming the core of a gene network; (2) the proteins encoded by these genes, performing structural, transportation, enzymatic, regulatory, and other functions; (3) pathways of signal transduction from cell membranes to cell nuclei, underlying either transcription activation or inhibition in response to external stimuli; (4) negative or positive feedbacks, stabilizing parameters of gene networks at a certain level or, on the contrary, shifting them from initial levels to transfer the system into another functional state; (5) low-molecular-weight components, such as hormones and other signal molecules; energy-storing components; various metabolites etc. switching the functions of gene network in response to external stimulus (Kolchanov et al., 2000).

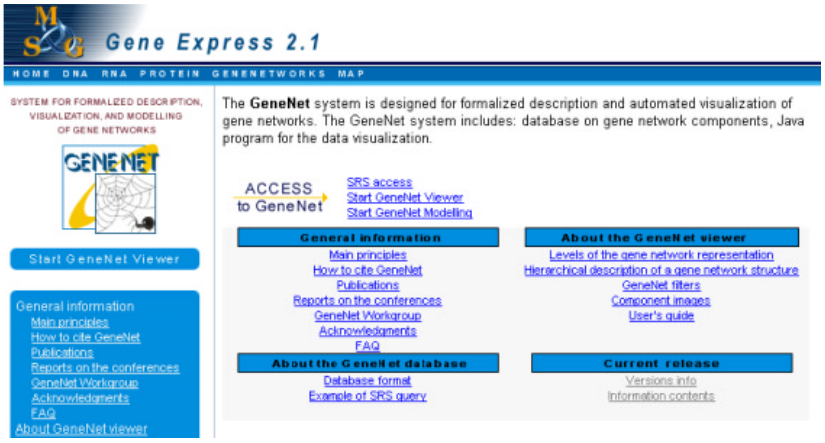


Fig. 1. Home page of the GeneNet system.

We undertook an attempt to compile any available information on gene networks in one database and in the context of a unified approach aiming to form the background for reconstructing an integrated pattern of the gene network function. For this purpose, the system GeneNet (Fig. 1) has been developed, as a part of the system GeneExpress (Kolchanov et al., 2000). GeneNet allows the formalized data on all the structural components of gene networks and a diversity of processes to be accumulated and the data compiled to be visualized as graphs (Kolpakov et al., 1998).

Methods and Algorithms

The technology for formalized description of structural components and elementary processes within gene networks was developed earlier (Kolpakov et al., 1998; Kolchanov et al., 2000). The program GeneNet Viewer (Kolpakov et al., 1997) is used to visualize the structure–function organization of gene networks; the specially developed program Data Input GUI (Kolpakov, Ananko, 1999), to input the data into GeneNet. All the tables constituting GeneNet are integrated using Sequence Retrieval System (SRS) v. 6.

Implementation and Results

So far, GeneNet comprises the descriptions of 25 gene networks under the 4 following sections: Lipid Metabolism, Endocrine Regulation, Morphogenesis, and Organism's Response to External Stimuli (Table 1).

Table 1. Sections of GeneNet.

GeneNet section	Name of gene network*	Number of elementary structures	Number of relationships
Lipid Metabolism	Cholesterol	30	34
	Cholesterol_MODEL	37	43
	Leptin (organism level)	125	89
Endocrine Regulation	Principal cell of CCD	32	34
	Steroidogenesis (adrenal cortex)	63	80
	Steroidogenesis (sex steroids)	66	78
	Thyroid system	89	110
Morphogenesis	Erythroid differentiation	103	98
	Germination (endosperm)	33	25
	LEA program	47	27
	Seed reserve mobilization (1): carbohydrates	24	34
	Seed reserve mobilization (2): lipids and phosphates	24	31
	Seed reserve mobilization (3): proteins	21	42
	Seed reserve mobilization (4): regulatory relationships	55	62
	Seed reserve mobilization (5): general diagram	59	59
	Seed reserve mobilization (organism level)	48	43
	Storage protein biosynthesis (dicots)	34	31
	Storage protein biosynthesis (monocots)	50	33
Organism's Response to External Stimuli	Antiviral response	67	53
	Macrophage activation (model)	125	124
	HSP70 autoregulation	32	37
	Plant–pathogen	81	65
	Heat shock response	103	114
	Thermotolerance	4	40
	REDOX regulation	52	64

*Find more details of gene network descriptions in this issue of BGRS-2002 Proceedings.

The current release 3.1 of GeneNet SRS-version as of April 1, 2002 comprises 14 interlinked tables and over 9000 entries (Table 2). The SRS version of the database GeneNet (Ananko et al., 2002) allows the complete list of the elementary structures for a gene network in question to be obtained and sorted according to structure types, species of the organisms, and localization in the cell. It also allows the reactions and regulatory events occurring in the gene network to be listed comprehensively. In addition, the user can find out the interactions of a protein of his (her) interest in all the available gene networks, the genes whose transcription is influenced by a transcriptional factor in question, etc.

The specially developed program GeneNet Viewer (Kolpakov et al., 1998), reading the data from the table GN_SCHEME to represent them as integrated graphs with interactive objects (Fig. 2) is used for visualizing the gene networks described.

The information compiled in the GeneNet database is also used for developing dynamic models of gene networks. The section GeneNet Modelling contains three dynamic models simulating three gene networks controlling lipid metabolism, erythropoietin-induced erythrocyte differentiation and maturation, and macrophage activation by LPS and IFN-γ. These models are detailed in the corresponding abstracts of this and previous issues of *BGRS Proceedings* (Ratushny et al., 2000; Ratushny et al., 2000; Nedosekina et al., 2002).

Table 2. Informational content of GeneNet database (SRS-version, release 3.1, April 1, 2002).

	Name of the SRS table	Contents	Number of entries
	GN_BIBLIOGRAPHY	References to the papers annotated	1377
	GN_CELL	Cells, tissues, or organs	367
	GN_COMPARTMENT	Compartments	135
	GN_EXPERT	GeneNet annotators	21
	GN_GENE	Genes	856
	GN_ORGANISM	Species	89
	GN_PROCESS	Input and output processes	194
	GN_PROTEIN	Proteins	1402
	GN_RELATION	Relationships between entities	1748
	GN_RNA	RNAs	280
	GN_SCHEME	Descriptions of gene networks	25
	GN_SCHEME_ENTITY	Entities (elementary structures)	1548
	GN_SCHEME_RELATION	Relationships in gene networks	1537
	GN_SUBSTANCE	Other substances	224

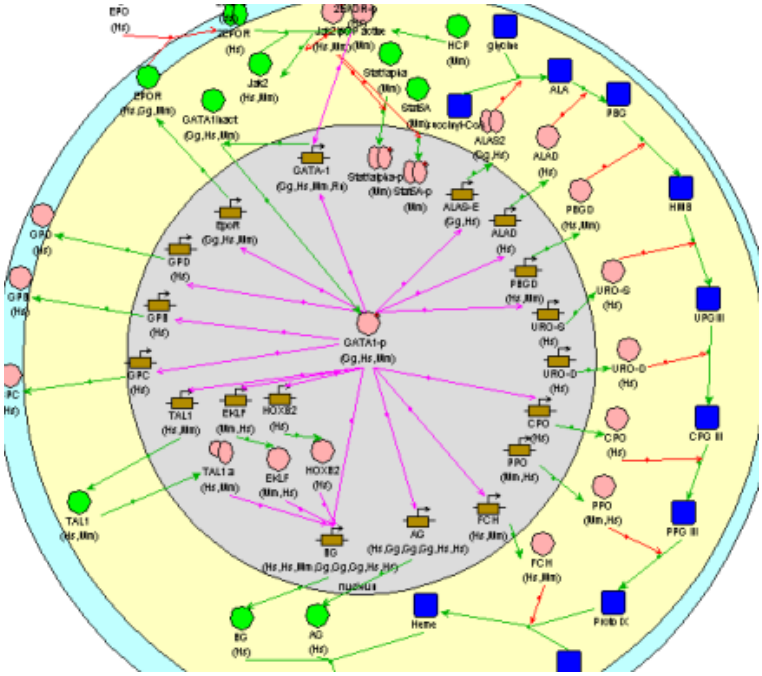


Fig. 2. A fragment of the integrated graphical scheme of the gene network controlling erythroid differentiation.

Discussion

The language specially developed for the GeneNet database allows any gene networks to be described, including symbiotic gene networks as well as signal transduction and metabolic pathways. Development of a new version of data editor using xml data format is now in progress. The new editor will be more convenient for the user and will allow most intricate gene networks to be described. The format used in GeneNet for describing elementary interactions will be modified, and the option to input both quantitative and qualitative information on process dynamics will be supplemented. A more intimate integration of the TRRD (Kolchanov et al., 2002) and GeneNet databases is also planned.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 00-07-90337, 00-04-49229, 00-04-49255, 01-07-90376, and 02-07-90359); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). Authors are grateful to I.V.Lokhova and L.V.Katokhina for bibliographical support; F.A.Kolpakov, A.Malinin, and E.Krestinin for development of software; D.A.Grigorovich for system administration; experts in biology A.V.Aksenovich, T.V.Busygina, T.N.Goryachkovsky, S.A.Grigoriev, N.S.Logvinenko, E.A.Nedosekina, and V.V.Suslov for annotating the information.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002). GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398-401.

2. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002). Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312-317.
3. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000). Gene networks. Mol. Biol. (Mosk.). 34:533-544.
4. 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 Yu.V., Titov I.I., Vishnevsky O.V., Orlov Yu.L, Valuev V.P., Ivanisenko V.A, Oshchepkov D.Yu., Omel'yanchuk N.A., Pozdnyakov M.A., Kosarev P.S., Goryachkovskaya T.N., Fokin O.N., Kalinichenko L.A., Kotlyarov Yu.V. (2000). Integrated system of gene expression regulation GeneExpress–2000. Proc. Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000). ICG, Novosibirsk, 1:12-18.
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. Kolpakov F.A., Ananko E.A. (1999). Interactive data input into the GeneNet database. Bioinformatics. 15,:13-714.
7. Nedosekina E.A., Ananko E.A., Likhoshvai V.A. (2002). Mathematical model of the gene network on macrophage activation under the action of IFN-γ and LPS. This issue (Proc. BGRS-2002).
8. Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. (2000). Mathematical model of cholesterol biosynthesis regulation in the cell. Proc. Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000). ICG, Novosibirsk. 1:199-202.
9. Ratushny A.V., Podkolodnaya O.A., Ananko E.A., Likhoshvai V.A. (2000). Mathematical model of erythroid cell differentiation regulation. Proc. Second International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2000). ICG, Novosibirsk. 1:203-206.

MOLECULAR-GENETICAL MECHANISMS OF ADIPOCYTE REGULATION: REPRESENTATION IN GENENET DATABASE

Proscura A.L., * Ignatieva E.V.

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

Key words: *database, GeneNet, lipid metabolism, adipocyte*

Resume

Motivation: Data accumulation and development of gene networks on the basis of the GeneNet technology that hold up a mirror to molecular-genetical mechanisms supporting functioning of various cells could be viewed as a supposition to development of the integrated gene network, which operates at the level of the whole organism. As known, obesity is a risk factor for many serious human diseases. Thereby, investigation of lipid metabolism regulation is of particular interest nowadays. The cells of adipose tissue (adipocytes) play a key role in lipid metabolism, the main function of these cells being the storage of energy in the form of triglycerides.
Results: By analyzing experimental information on the mechanisms of regulation of gene expression in adipocytes, as well as on the data on metabolic reactions, we have designed the first release of the adipocyte gene network within the frames of the GeneNet database. In this work, we present description of the gene network and its logical analysis.
Availability: the GeneNet system is available at http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/.

Introduction

The lipid metabolism system provides vital functions of an organism, including regulation of lipid assimilation from food, lipid transportation via the blood flow, utilization of lipids in a cell, synthesis *de novo*, destruction and excretion from the organism. Disruption of lipid metabolism is a causative reason of many serious human diseases: atherosclerosis, ischemic cardiomyopathy, etc. Due to this reasoning, this system is an object of intense interest (Schmitz et al., 1998). The database GeneNet (Ananko, this issue) contains the section «Lipid metabolism», which includes information about several subsystems of lipid metabolism (Fig. 1). Among these data are those referring to two processes, cellular cholesterol regulation (the diagram «Cholesterol») and body weight regulation with participance of leptin (the diagram «Leptin» («organism level»)), as well as the data on molecular-genetic mechanisms of functioning of an adipocyte («Adipocyte1»). Adipocytes play a key role in lipid metabolism regulation. Its main function is the storage of energy in a form of triglycerides. Triglycerides are synthesized from glycerol and saturated or unsaturated fatty acids (Fig. 2, a), e.g., palmitate (Fig. 2, b). The rate of this process depends upon food consumption and it is regulated by hormones (insulin, glucocorticoids). In its turn, an adipocyte provides secretion of the hormone leptin (the product of the ob gene), which influences the hypothalamic areas important in the control of food intake (Friedman, Halaas, 1998).

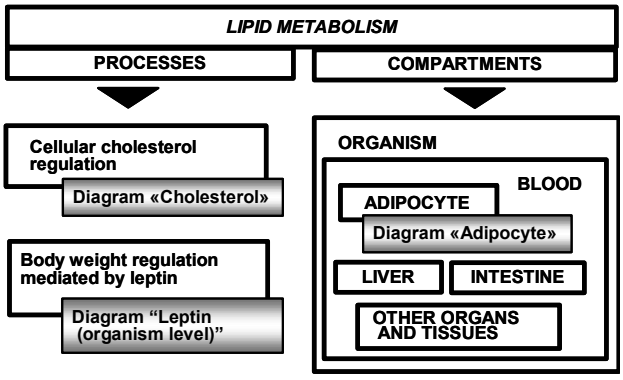


Fig. 1. Schematic representation of the «Lipid metabolism» section in the GeneNet database.

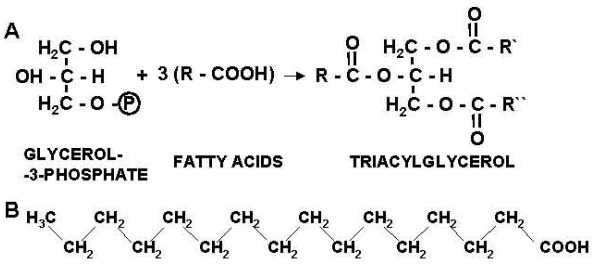


Fig. 2. a) General scheme of triglyceride synthesis.b) Structure of palmitic acid (palmitate).

In the first release of the gene network on adipocyte, the data are stored on 16 genes, 13 mRNAs, 30 proteins, 100 relations between the objects, which were accumulated in the database in accordance with analysis of 34 experimental publications.

Methods

Input of information into the GeneNet system was made via the Internet by using the system of interactive data submission (Kolpakov, Ananko, 1999).

Results

In the first release of the gene network on adipocyte regulation, five biochemical processes are included: input and utilization of glucose and fatty acids, biosynthesis of fatty acids, triglycerids, and cholesterol.

Fatty acids synthesis is produced under participance of enzymes, ACC (acetyl-CoA-carboxylase) and FAS (fatty acid synthase). First, acetyl-CoA in carboxylation reaction catalyzed by ACC turns into malonyl-CoA (Fig. 3, 1a). Then, by means of the enzyme FAS, the cycle of subsequent modification from acetyl-CoA and malonyl-CoA till acyl-enzyme (acyl-enzyme is not shown at the scheme) is carried out. The cycle is repeated 6 times, until saturated fatty acid, 16-carbon palmitate is not formed (Fig. 3, 1b). Next, transformation of palmitate into unsaturated fatty acid (palmitoleoil-CoA) is possible under the action of enzymes, acetyl-CoA synthetase (ACS) and stearoyl-CoA desaturase (SCD). During this process, palmitate first turns into palmitoyl-CoA (enzyme ACS) (Fig. 3, 1c), then palmitoyl-CoA enters reaction of desaturation and double bonds are formed (enzyme SCD) (Fig. 3, 1d).

Synthesis of triglycerols is produced from acyl-CoA-derivatives of fatty acids and glycerol-3-phosphate. In this reaction, acyl-CoA-derivatives of saturated and unsaturated fatty acids, both of endogeneous origin (i.e., those synthesized in a cell, for example, palmitoyl-CoA, palmitoleoyl-CoA) (Fig. 3, 2b) and of exogenous origin (those incoming outside from the cell, e.g., acyl-CoA derivatives of fatty acids) (Fig. 3, 2a) may participate.

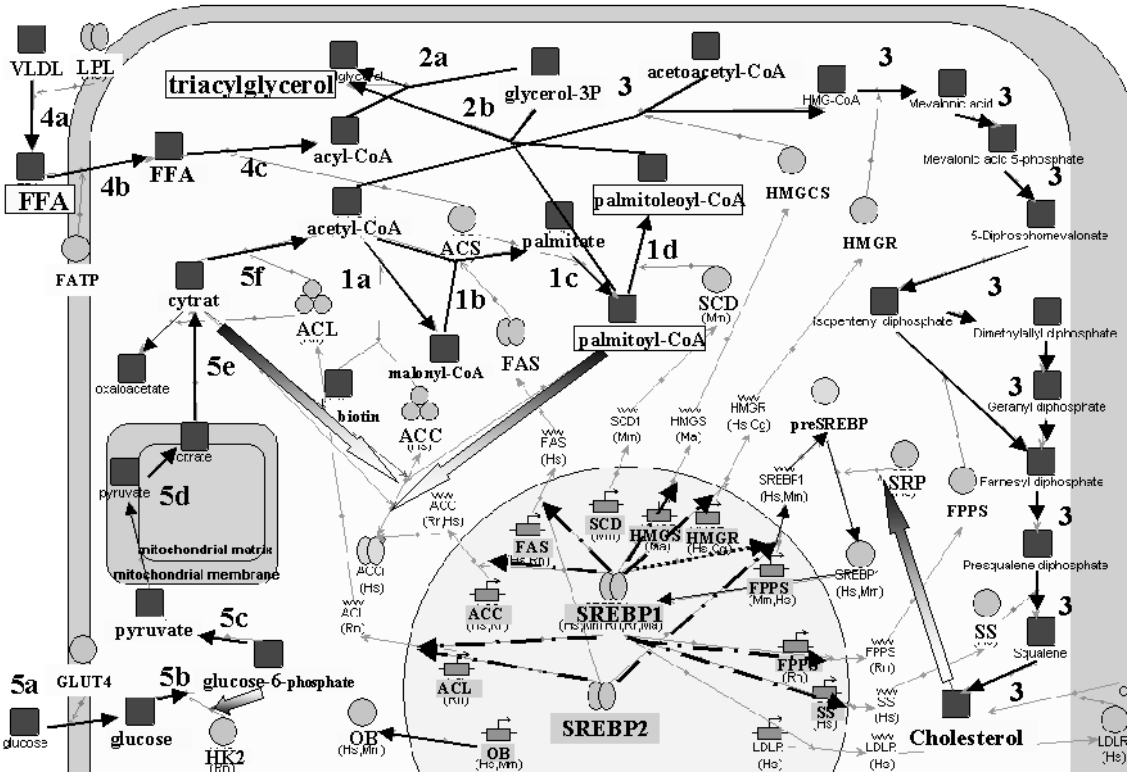


Fig. 3. Graphical representation of information from the section "Adipocyte1" of the GeneNet database.Denotations. By solid arrows, metabolic pathways are marked: 1, (a-d, fatty acids synthesis; 2, (a-b) synthesis of triacylglycerols; 3, cholesterol biosynthesis; 4, (a-c) input and utilization of fatty acids, 5, (a-f) input and utilization of glucose. By chain-line arrows, the influence of transcription factors on the gene expression is shown. Dotted arrow marks autoregulation of the SREBP gene expression . Large black-white arrows denote regulation of activity of enzymes by initial substrates or by the products of biochemical reactions.

Cholesterol biosynthesis proceeds by using as initial substances of acetyl-CoA and acetoacetyl-CoA. This process is catalyzed by several enzymes, including HMG-CoA synthase (HMGCS), HMG-CoA reductase (HMGCR), farnesyl diphosphate synthase (FPPS), and squalene synthase (SS) (Fig. 3, 3).

Input and utilization of fatty acids. The enzyme lipase (LPL) hydrolyzes triacylglycerols, which are transported to the cell membrane of adipocytes by chylomicrons and VLDL (very low density lipoproteins), with fatty acids (FFA) release (Fig. 3, 4a). FFA are transported into the cell by the fatty acid transport protein (FATP) (Fig. 3, 4b), then, under the action of the ACS enzyme, turn into acyl derivative, acyl-CoA (Fig. 3, 4c).

Input and utilization of glucose. The passage of glucose through the cell membrane proceeds via the transport protein GLUT4 (Fig. 3, 5a). Then glucose is utilized in glycolysis. Hereby, it is first phosphorylated by hexokinase II (HK2) into glucose-6-phosphate (Fig. 3, 5b) and, then, via a seria of reactions, it turns into pyruvate (Fig. 3, 5c). Pyruvate incomes to mitochondria, where it turns into citrate in the Krebs cycle (fig.3, 5d). Citrate moves into outer-mitochondrial compartment (Fig. 3, 5e), where ATP citrate-lyase (ACL), with the uptake of CoA and ATP, catalyzes its disintegration into acetyl-CoA and oxaloacetate (Fig. 3, 5f). Acetyl-CoA is an initial substrate for reactions of fatty acids biosynthesis, in particular, of palmitic acid biosynthesis.

Regulation of this gene network is controlled by several ways:

1) Intensity of the triglycerol and cholesterol biosyntheses depends upon input into the cell of glucose and fatty acids.

a) citrate (the product of desintegration of glucose) is a source of the acetyl-CoA. In its turn, acetyl-CoA is one of initial substances needed for cholesterol and triglycerol biosynthesis;

b) acyl-CoA is synthesized on the basis of free fatty acids incoming into the cell. It is one of initial substances for the triglycerol biosynthesis.

2) Both metabolic pathways (fatty acid and cholesterol biosynthesis) use as an initial substance one and the same chemical agent, acetyl-CoA.

3) Activity of a seria of enzymes is regulated by initial substrates or by the products of biochemical reactions, catalyzed by these enzymes (they are marked in fig.3 by large black-and-white arrows).

a) Citrate (or the molecule, from which acetyl-CoA is formed) activates transition of the enzyme, ACC, using acetyl-CoA as a substrate, into the active dephosphorylated state.

b) Palmitoyl-CoA, formed from acetyl-CoA and malonyl-CoA under the action of two enzymes, ACC and FAS, suppresses activity of ACC, thus, accelerating its transition into inactive state.

c) Glucose-6-phosphate, formed under the action of hexokinase, inhibits activity of this enzyme.

d) Cholesterol, produced in the cell, inhibits expression of the enzymes catalyzing its biosynthesis (HMGCS, HMGCR, FPPS, and SS) through suppressing activity of sterol regulated protease (SRP), which splits the preSREBP protein, inactive precursor of the SREBP transcription factor (sterol regulatory element-binding protein). SREBP stimulates expression of the genes: HMGCS, HMGR, FPPS, and SS. If the cholesterol level is high, then SRP activity decreases, then concentration of the active SREBP protein falls, thus reducing transcription of genes controlling cholesterol biosynthesis (HMGCS, HMGR, FPPS, and SS) (Fig. 3, chain line).

4) Transcription factors that are expressed in adipocytes regulate their own transcription (autoregulation). For example, transcription factor, SREBP1, activates transcription of the SREBP1 gene (Fig. 3, dotted line).

Discussion

The gene network presented enables to view the processes of lipid metabolism both in norm and under pathology. As known from literature data, in adipose tissue of obese rats (fa/fa Zucker rats with mutation of the fa gene), an expression of the FAS gene is accelerated, this gene being a key element of the fatty acids biosynthesis. As shown, the FAS gene has a negative regulatory region (fa-responsive region), which functions in adipose cells of lean rats and loses its activity in the cells of obese rats. In adipocytes of the fa/fa Zucker rats, concentration of the SREBP2 transcription factor is increased in comparison to the norm. Due to supposition of the authors of publication, increased concentration of SREBP2 prevents the negative effect of the fa-responsive region on the FAS gene transcription. As a consequence, extremely high production of the FAS enzyme causes enhanced accumulation of fat in these animals (Boizard et al., 1998).

Conclusion

Currently, the GeneNet database contains several gene networks accumulating the data on molecular and genetic mechanisms of lipid metabolism regulation (Fig. 1). The diagrams «Cholesterol» and «Leptin» illustrate the data on the processes regulating the cholesterol cellular level and the body weight mediated by leptin, respectively. The gene network «Adipocyte1» represented in a work given includes the data on peculiarities of the lipid metabolism in adipose tissue, in particular, the data about the main function of adipocytes, triglycerole biosynthesis. Subsequent releases of the gene network on adipocyte cell will be supplemented by novel data on biochemical pathways and on the genes regulating them, on endocrine control of gene expression in adipocytes, and on the mechanisms of adipocyte proliferation. The gene network «Adipocyte1» is the first in this section on the lipid metabolism, which includes the data on functional peculiarities of a particular cell type. Further, we plan to develop the gene network on regulation of lipid metabolism in liver cells, in gastric sells, and on lipid transport in blood. Integration of the local gene networks will provide possibility to reconstruct the gene network of lipid metabolism at the level of an organism and, further, to develop a mathematical model of this gene network. It is supposed also to accumulate quantitative data about dynamics of gene network functioning in order to verify mathematical models on the basis of the technology suggested by V.A.Likhoshvai (Likhoshvai et al., this issue).

Acknowledgements

The authors are grateful to Professor N.A.Kolchanov for fruitful discussions, to E.A.Ananko for help and consultations during the work with the GeneNet system, to I.V.Lokhova and L.V.Katokhina for bibliographical support. The work was

supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, 00-04-49255), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65, 66), National Institutes of Health USA (grant № 2 R01-HG-01539-04A2), The Department of Energy USA (grant № 535228 CFDA 81.049).

References

1. Ananko E.A. (2002) GeneNet system, its status in 2002. (This issue).
2. Boizard M., Le Liepvre X., Lemarchand P., Foufelle F., Ferre P., Dugail I. (1998) Obesity-related overexpression of fatty-acid synthase gene in adipose tissue involves sterol regulatory element-binding protein transcription factors. J. Biol. Chem. 273, 29164-29171.
3. Friedman J.M., Halaas J.L. (1998) Leptin and the regulation of body weight in mammals. Nature. 395, 763-770.
4. Kolpakov F.A., Ananko E.A. (1999) Interactive data input into the GeneNet database. Bioinformatics. 15, 713-714.
5. 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.
6. Likhoshvai V.A., Latypov A.F., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. (2002) Technology of using experimental data for verification of models of gene network operation dynamics. (This issue).

GENE NETWORK OF GLUTATHIONE HOMEOSTASIS: A RESPONSE TO OXIDATIVE STRESS

**** Kudryavtseva A.N., Stepanenko I.L.***

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

Key words: *gene networks, signal transduction pathways, metabolic pathways, glutathione, hydrogen peroxide, oxidative stress*

Resume

Motivation: Oxidants and antioxidants regulate the redox balance in a cell and, hence, antioxidant manipulation might be a potential way to control gene expression.
Results: Using the technology of the GeneNet database, a gene network regulating glutathione homeostasis in a cell was constructed. Glutathione is involved into many cell processes, from antioxidant protection to proliferation modulation. Since imbalance of homeostasis can lead to various pathologies, a mathematical model of the gene network operation can be used for medical purposes, particularly, in cancer therapy.
Availability: http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet.

Introduction

Reactive oxygen species (ROS) are formed in a cell during cell metabolism and under the action of various growth factors and cytokines. ROS regulate different physiological functions, mainly, those involved in signal transduction pathways. They also regulate gene expression via cysteine residues in the DNA-binding domains of transcription factors and inactivate cysteines at the catalytic sites of protein phosphatases, thus modulating the kinase activity (Gabbita et al., 2000). Oxidants and antioxidants regulate the redox balance in a cell and, hence, antioxidant manipulation might be a potential way to control gene expression.
Glutathione and thioredoxin are two main antioxidants involved into redox regulation in cells. Glutathione (GSH), the thiol tripeptide-antioxidant γ-Glu-Cys-Gly, can be in the oxidized (GSSG) or reduced forms. Conversion of one form to another is catalyzed by the enzymes glutathione reductase (GR) and glutathione peroxidase (GPX1, GPX2):
2 GSH + H(2)O(2) = GSSG + 2H(2)O.
Acting as a buffer system, the ratio of these two forms maintains a redox potential in different cell compartments at a certain level. Glutathione performs many functions in a cell, from antioxidant protection to proliferation modulation (Lu, 1999). The glutathione level is low in patients with type II diabetes, cataract, neurological disorder, AIDS, and hepatitis C (Rahman, MacNee, 2000). A connection between the polymorphism of the enzymes involved in glutathione detoxification, synthesis, and transportation with drug resistance was discovered as well as the probability of successful cancer chemotherapy (Tsuchida, Sato, 1992).

Methods

We used a computer technology employed in the gene network database GeneNet (Ananko et al., 2002). Experimental data on the regulation of glutathione homeostasis collected by annotating scientific literature was input into the database using the Data Input application program. The data visualization program GeneNet Viewer displays formalized data on the structure–function organization of the gene network of glutathione homeostasis.

Results

Activation of the gene network regulating glutathione homeostasis in a cell. An imbalance in the system glutathione–active oxygen results in the activation of a kinase cascade and gene expression, regulating glutathione homeostasis. The key transcription factor Nrf2 (nuclear factor erythroid 2-related factor 2) in an inactive form is localized to the cytoplasm in complex with the anchor protein Keap1. During the formation of ROS in the cell, Nrf2 translocated from the cytoplasm to the nucleus initiates gene expression of enzymes involved in glutathione synthesis and enzymes regulating glutathione redox status and the transport of limiting cysteine amino acid. One of the pathways is phosphorylation of the Nrf2/Keap1 complex by protein kinase C (PKC), which promotes Nrf2 transport into the nucleus and activates gene transcription. The response rate depends on the velocity and duration of Nrf2 transport, which, in turn, depends directly on the concentration of hydrogen peroxide. The transcription factor Nrf2, a homodimer, cannot bind to DNA but activates target genes in complex with small proteins Maf and Jun. The transcription factor Nrf1, also involved in the gene network regulation, activates the transcription of GCS genes and several glutathione transferases but binds to DNA with a low affinity in comparison with Nrf2.

Inhibition of gene network functioning by varying the balance of transcription factors. In response to oxidative stress, gene expression of the transcription factors Nrf2, c-Jun (JunB), c-Fos, MafG, and Fra-1 increases. The response to oxidative stress fails if the ratio of transcription factors in the nucleus changes. However, when the transport of proteins from the nucleus (disturbed by hydrogen peroxide) is restored, Nrf2 is transported to the cytoplasm, where it binds to the Keap1 protein, whose expression increases in response to oxidative stress. The expression of Maf and Fra-1 genes also increases in response to oxidative stress. Homodimers Maf/Maf and heterodimers Fra-1/Jun bind to ARE and the Nrf2 binding sites and repress the genes regulating glutathione synthesis.

Inhibition of gene network function by glutathione. Activation of the gene network is inhibited by glutathione at several levels. Glutathione is synthesized by two enzymes: gamma-glutamylcysteine synthetase (GCS) and glutathione synthetase (GS). GCS is a key, rate-limiting biosynthesis enzyme consisting of two subunits, namely regulatory and catalytic, encoded by the GCSL and GCSH genes. The enzyme in the synthesis of GCSh is inhibited by its product, glutathione. First, intersubunit binding depends largely on the oxidation–reduction potential of the cell. Note that in a reduced state, the enzyme activity is lower. Second, the enzyme is inhibited due to direct binding of glutathione to a heavy catalytic subunit. Glutathione inhibits activation of PKC kinases.

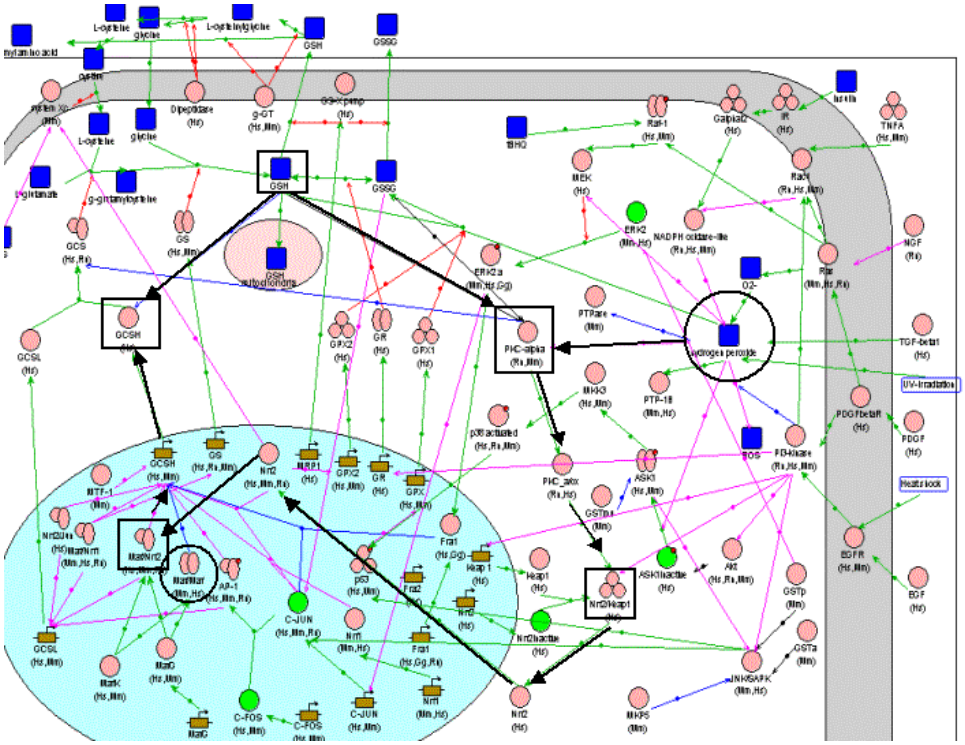


Fig. Gene network regulating glutathione homeostasis in response to hydrogen peroxide. Hydrogen peroxide treatment of cells induces activation of kinases. Phosphorylation of the Nrf2/Keap1 complex by PKC promotes transport of Nrf2 into the nucleus and activation of the gene network. In response to oxidative stress, the expression of Maf and Fra-1 genes also increases. The balance of transcription factors Nrf2, Maf, and Jun regulates gene expression of glutathione synthesis. The activity of glutathione synthesis enzyme GCSh depending on the redox status of the cell is inhibited by glutathione binding.

Discussion

Specific of the organization of gene networks is the ability to self-regulate due to closed regulatory circuits with positive and negative feedbacks (Kolchanov et al., 2000). This gene network belongs to typical gene networks regulating the maintenance of homeostasis in a cell. Regulatory circuits with negative feedbacks provide maintenance of the gene network parameters at a certain level. Special characteristics of this gene network are determined by a key regulating transcription factor. Nrf2 belongs to the CNC-bZIP subfamily of the bZIP transcription factors, including Jun, Fos, ATF/CREB, and Maf factors. Unlike other CNC-bZIP proteins containing a cap`n`Collar (CNC) motif, Nrf2 is unable to form homodimers. Therefore, for binding to DNA and transactivating, Nrf2 forms complexes with other bZIP transcription factors and small proteins Maf and Jun. Expression of Nrf2-regulated genes encoding proteins that play a key role in adaptive response to oxidative stress (heme oxygenase, NAD(P)H:quinone oxidoreductase, γ-glutamylcysteine synthetase, glutathione S-transferase) depends on the balance of Nrf2-binding transcription factors in the nucleus.

Thus, this gene network has two levels of regulation. One level includes direct inhibition of the glutathione synthesis enzyme by glutathione and activation of the gene network via PKC, a regulatory circuit with a negative feedback. Another level of regulation is determined by the ratio of the transcription factors in the nucleus under oxidative stress. The

composition of the Nrf2-containing complex and variation in the level of gene expression encoding these transcription factors modulate functioning of the gene network, depending on the cell type and inductor.

Being an important element in the system of cell detoxification, glutathione conjugates to xenobiotics and their metabolites by the entire class of enzymes glutathione transferases (GSTp and GSTmu). Then, glutathione conjugates (oxidized glutathione under oxidative stress and reduced glutathione during transport from the liver to blood and bile) are transported by the GS-X pump beyond cell borders. Glutathione and the conjugates are degraded by the membrane-bound enzymes gamma-glutamyltranspeptidase (g-GT) and dipeptidase only outside the cell. After hydrolysis, amino acids return to the cell for glutathione synthesis.

Prospects of modelling this gene network is a direct approach to the problem of drug resistance of cancer cell lines, which is related to increased activity of glutathione transferases, active transport of glutathione conjugates and drugs, and glutathione synthesis. The dynamic database (Likhoshvai et al., 2002) stores quantitative data on the dynamics of the gene network regulating glutathione homeostasis, which will be further used to verify models of gene networks.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002) GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398–401.
2. Gabbita S.P., Robinson K.A., Stewart C.A., Floyd R.A., Hensley K. (2000) Redox regulatory mechanisms of cellular signal transduction. Arch. Biochem. Biophys. 376:1–13.
3. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaia O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000) Gene networks. Mol. Biol. (Mosk.). 34: 533–544.
4. Likhoshvai V.A., Latypov A.F., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. (2002) Technology of using experimental data for verification of models of gene network operation dynamics. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
5. Lu S.C. (1999). Regulation of hepatic glutathione synthesis: current concepts and controversies. FASEB J. 13:1169–83.
6. Rahman I., MacNee W. (2000) Regulation of redox glutathione levels and gene transcription in lung inflammation: therapeutic approaches. Free Radic. Biol. Med. 28:1405–20.
7. Tsuchida S., Sato K. (1992) Glutathione transferases and cancer. Crit. Rev. Biochem. Mol. Biol. 27:337–384.

MOLECULAR GENETIC MECHANISMS REGULATING THE THYROID SYSTEM: DESCRIPTION IN THE TRRD AND GENENET DATABASES

Suslov V.V. [*]***, Ignat'eva E.V.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[*]Corresponding author

Key words: *databases, ES-TRRD, thyroid system, transcription regulation, gene networks, GeneNet*

Resume

Motivation: Thyroid disorders unless caused by iodine deficiency result from impairments of certain regulatory links in the gene network. Our understanding of the gene network as a self-regulating system allows us to model these disorders and correct them using undamaged regulatory links. Data collection and analysis preceding modelling are more convenient to perform using specially designed databases with visualization facility.
Results: The TRRD database stores formalized data on transcription regulation of 19 genes of the thyroid system. The data about the function of gene network regulating the thyroid system are stored in formal state in GeneNet database. A logic analysis of the gene network of the thyroid system was performed in order to detect the key regulatory links for subsequent mathematical modelling.
Availability: ES-TRRD is available in the Internet at http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd; and GeneNet is available at http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet.

Introduction

The thyroid system regulates a basic metabolic level and, consequently, (a) works in cooperation with other systems in a body (insulin regulatory system, somatotropic function, etc.) and (b) affects the formation of many organs (in particular, brain) and tissues with a high level of metabolic activity. Thus, a dysfunction of this system influences the entire body. With account for susceptibility of the thyroid gland to autoimmune diseases, collection and storage of experimental data on the thyroid system in an electronic form are of great importance. Systematization and computer analysis of these data will yield mathematical models describing normal and pathologic processes.
TRRD (Transcription Regulatory Regions Database), developed at the Institute of Cytology and Genetics, Siberian Branch of the Russian Academy of Sciences (Kolchanov et al., 2002), accumulates the data on structure–function organization of transcriptional regulatory regions of genes and the specific features of gene expression. The GeneNet database (Ananko et al., 2002) gives a visual representation of the structure of gene networks. Accumulation of data in the TRRD and GeneNet databases facilitates computer analysis and mathematical modelling of physiological processes. This paper describes the information content of the subunits of these databases that deal with the regulation of the thyroid system.

Methods

Data on regulation of gene transcription in the TRRD database format were accumulated in ES-TRRD, the section *Thyroid System* (Kolchanov et al., 2002). Data on the structure of the gene network of the thyroid system were input into the GeneNet database (Ananko et al., 2002) using the system of interactive data input available in the Internet (Ananko et al., 1999).

Results

TRRD database. The section ES-TRRD (Endocrine System–Transcription Regulatory Regions Database) stores information on regulation of transcription of 19 thyroid genes. We described genes of hormones and their precursors, enzyme genes, and genes of transcription factors and their receptors. ES-TRRD describes 22 regulatory regions (promoter, enhancer, etc.) containing 76 transcription factor binding sites. This information was collected by analyzing 60 scientific publications.
GeneNet database. The GeneNet section *Thyroid System* describes hypothalamic–pituitary–thyroid interactions, a central fragment of the gene network of the thyroid system. In addition, the database contains data on (a) cytokine regulation of genes expressed in thyrocytes, (b) thyrotropin (TSH) and endogenous somatostatin regulations of thyrocyte proliferation, and (c) interactions of the thyroid and somatotropic systems with the epiphysis. The section *Thyroid System* includes the data on 87 genes, 139 proteins, 81 RNAs, 11 processes, and 410 links between the objects. The information was collected through analyzing 130 scientific papers.

Logical analysis of the gene network. The subnetwork of gene regulation and synthesis of thyroid hormones is described most completely (Fig. 1). This subnetwork includes two interacting regulation mechanisms: the first one is a relatively autonomous, cellular mechanism depending on the TTF-1 transcription factor; the second mechanism is hypothalamic–pituitary depending on TSH (thyroid stimulating hormone).

TTF-1-dependent regulation mechanism. TTF-1 (thyroid transcription factor-1) plays an important role in differentiation of the thyroid gland. Its expression is activated by HOX proteins, the products of homeotic hox-genes, during early embryogenesis (Fig. 1).

1) TTF-1 gene regulation. The promoter of TTF-1 gene contains binding sites for its product, TTF-1 protein. Thus, the ***first positive feedback (PF) loop*** is formed (Fig. 2), leading to self-activation of the system.

2) The TTF-1 protein produced activates transcription of the cassette of thyroid hormone synthesis genes, namely, thyroglobulin gene (TG), thyroperoxidase (TPO) gene, and sodium/iodide symporter (NIS). In addition, TTF-1 activates transcription of thyrotropin receptor gene (TSHR). Thyrotropin (TSH) is a pituitary hormone regulating synthesis of thyroid hormones. TSHR is expressed on the basal membrane of the thyrocyte, which becomes competent (able to receive signals) towards the hypothalamic–pituitary system (Fig. 2).

3) The product of the NIS gene pumps iodine from blood through the basal membrane into the thyrocyte cytoplasm. The enzyme thyroperoxidase, the TPO gene product, is accumulated on the apical membrane (follicular lumen membrane) of thyrocyte and especially in vesicles surrounding the membrane. The TG gene product—a precursor of thyroid hormones, thyroglobulin—is accumulated in the follicular lumen (Fig. 1).

4) Accumulation of thyroglobulin in the follicular lumen activates PDS gene (the Pendred's syndrome gene) transcription. The mechanism of PDS gene activation has not been yet detected in detail. The PDS gene product, the protein pendrine, provides the delivery of iodine from the cytoplasm to the follicular lumen. Iodination of tyrosine residues in thyroglobulin is activated by thyroperoxidase. Highly iodinated thyroglobulin is unable to activate the PDS gene. Due to this, iodine transport through the apical membrane decreases, and ***the first negative feedback (NF) loop*** forms (Figs. 1, 2), which regulates the iodine influx to the follicular lumen, and, hence, synthesis of thyroid hormone.

5) On the other hand, highly iodinated thyroglobulin can bind to asialoglycoprotein receptors (ASPGR) localized to the apical membrane. Consequently, inhibition of the NF I factor results in inhibition of the TTF-1 synthesis, and, hence, inhibition of the synthesis genes of the thyroid hormones (TG, TPO, and NIS). Thus, the ***second NF loop*** is formed (Figs. 1, 2).

6) In addition, the TSHR transcription is inhibited, and thyrocyte becomes refractory (Figs. 1, 2). Thus, TTF-1-dependent regulation allows thyrocyte to start transmembrane iodine transport and synthesis of thyroid hormones as well as regulate both processes and the degree of competence to signals from the hypothalamic–pituitary system.

TSH-dependent regulation. Adenohypophysial thyrotropin (TSH), the main external regulator of thyroid hormone synthesis, consists of two subunits: an α-subunit common for glycoprotein hormones and a specific β-subunit.

1) Via the receptor on the basal membrane, TSH in a cAMP-dependent manner activates (a) transcription of the transcription factor TTF-2, which activates transcription of the TPO and TG genes; (b) release of the previously produced TPO enzyme from vesicles; (c) transcription of NIS gene via an insufficiently studied factor NTF-1 (Na+/I- symporter TSH responsive factor-1); and (d) transcription of Pax-8 gene, whose product of this gene activates genes TG, TPO, and NIS and is very important for the interaction of the TTF-1-dependent and TSH-dependent regulatory pathways (Fig. 1, 2).

2) Finally, TSH activates the transcription of its receptor TSHR via CREB protein (involving TTF-1), thus forming the ***second PF loop*** (Fig. 2). This leads to a quick (1–2 hr) production of TSHR in thyrocyte (the thyrocyte competence increases quickly; Figs. 1, 2).

3) However, a long TSH stimulation (> = 4 hr) results in the formation of the ***third NF loop*** (Fig. 2), which inhibits TSHR synthesis and leads to thyrocyte refractoriness. This loop is cAMP-dependently implemented by activation of ICER protein (a one of products of the CREM gene) expression and/or via CREB and SSB proteins (Figs. 1, 2).

4) In addition, thyroid hormones (for simplicity, the diagram shows only T3, triiodothyronine) decrease the thyrocyte competence, inhibiting the TSHR gene transcription via their receptors. In this case, the ***fourth NF loop*** is obvious (Fig. 1, 2). The same hormones inhibit in a similar manner the synthesis of TSH α- and β-subunits in adenohypophysis thyrotrophs and the synthesis of prepro-TSH-releasing hormone in hypothalamus. In this case, the ***fifth and sixth NF loops*** are observed (Figs. 1, 2). Thus, the TSH-dependent regulation leads to homeostasis of concentration of thyroid hormones in blood at different levels.

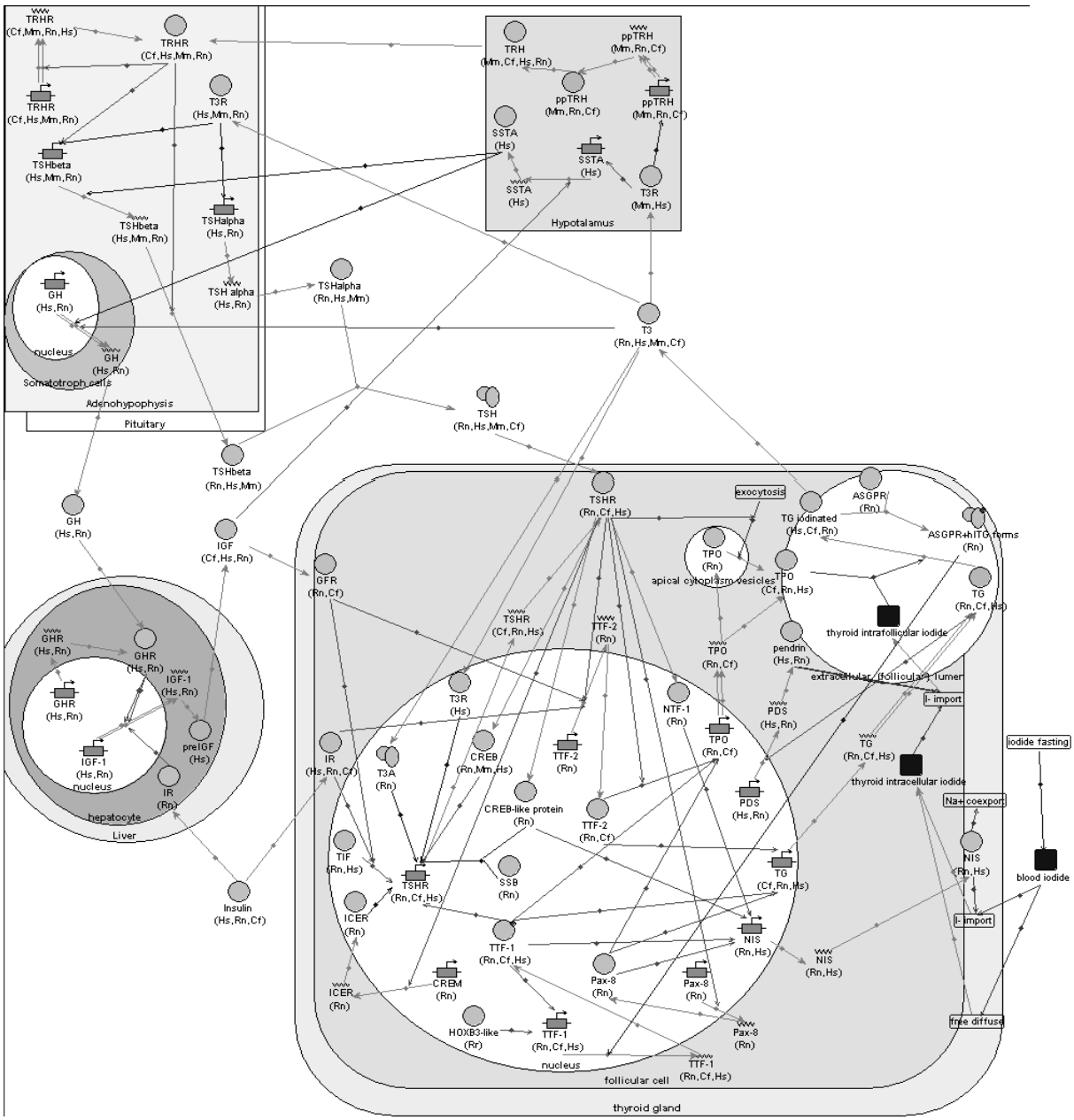


Fig. 1. Graphical representation of regulation of gene expression of the thyroid hormone synthesis in GeneNet. References to literature see in the databases at http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet.

Discussion

Thus, the gene network of the thyroid system belongs to the second type of gene networks that ensures homeostasis to numerous negative feedbacks. However, loops with both positive and negative feedbacks are present at the receptor level and the level of the transcription factor TTF-1. This is explained by the necessity of rapid and fine adjustment of the processes at these levels, and, in the case of TTF-1, by its involvement into the ontogenesis of the thyroid gland. For mathematical modelling of the thyroid system within the gene network, it is necessary to specify the following issues: (a) the level of intracellular regulation in the hypothalamus and pituitary gland; (b) the mechanisms of cytokine action on the thyrocyte; c) the TSH-dependent control of the genes regulating thyrocyte proliferation; and (d) the set of genes vital for thyrocyte that control the H_2O_2 level and redox regulation of thyrocyte genes. Since specific of the thyroid system is its cooperative interaction with other systems, the *Thyroid System* gene network can be used as a model object for virtual integration of gene networks.

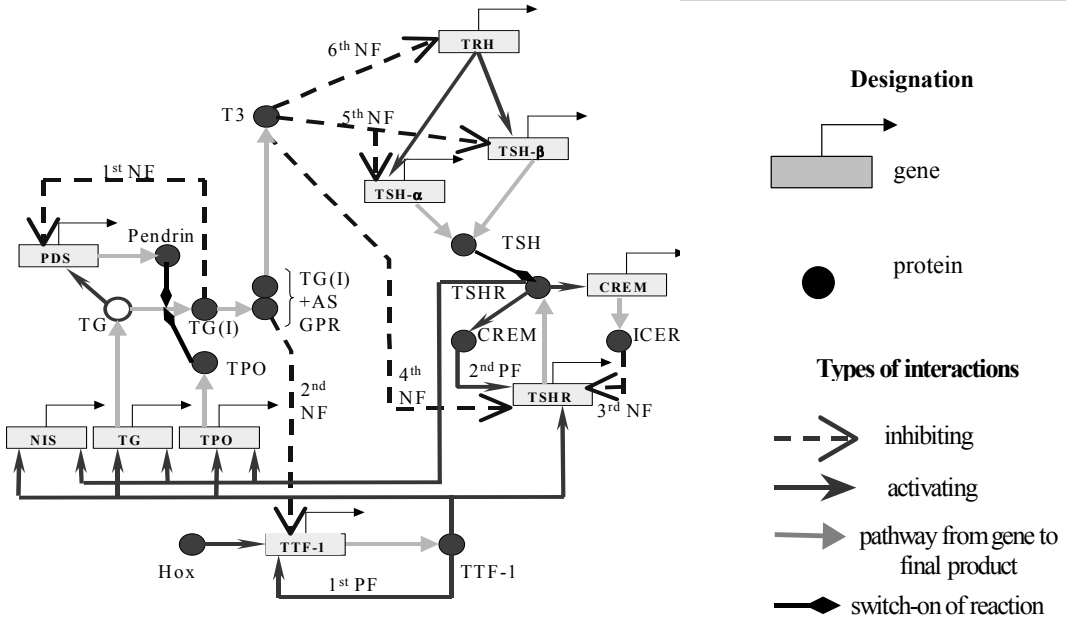


Fig. 2. Logical scheme of the regulation of expression of the genes controlling the synthesis of thyroid hormones (all abbreviations are given in the text, references to literature see in the databases at http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet).

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90359, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 65); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank Prof. N.A. Kolchanov for his helpful discussions, I.V.Lokhova and L.V.Katokhina for their assistance with bibliography, and I.V.Filippova for translation into English.

References

1. Kolchanov N.A., Ignatieva E.V., Ananko E.A., Podkolodnaya O.A., Stepanenko I.L., Merkulova T.I., Pozdnyakov M.A., Podkolodny N.L., Naumochkin A.N., Romashchenko A.G. (2002) Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucl. Acids Res. 30:312–317.
2. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002) GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398–401.
3. Ananko E.A., Kolpakov F.A. Interactive data input into the GeneNet database (1999). Bioinformatics. 15:713–714.
4. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000). Gene networks. Mol. Biol. (Mosk.). 34:533–544.

INTERPRETATION OF GENE NETWORKS IN THE CONTEXT OF ANOKHIN'S THEORY OF FUNCTIONAL SYSTEMS

**** Suslov V.V.*** [1]***, Vityaev E.E., Ignatieva E.V.***

Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[1] Institute of Mathematics, SB RAS, Novosibirsk, 630090, Russia, e-mail: valya@bionet.nsc.ru; vityaev@math.nsc.ru
*Corresponding author

Key words: *gene networks (GN), control, functional systems, GeneNet, thyroid system, thyroid hormones*

Resume

Motivation: Gene networks (GNs) are intricate systems responding constantly to changes in the environments by targeted rearrangements of their structure. Consequently, GNs optimize their operation in accordance with the changed conditions, that is, reach the goal. The set of goals and methods of their implementation is formed during the evolution of particular GN. The general patterns that determine the GN strategy of decision making at the molecular genetic level are yet vague. The theory of functional systems (TFS), developed by P.K. Anokhin, suggests an approach to interpreting the principles of GN organization and function. In this work, the operation of thyroid system GN, contained in the GeneNet database, is interpreted in the context of TFS.
Results: The thyroid system GN is logically analyzed. The detected hierarchy of results is compared to the structure–function organization of this GN. It is demonstrated how functional systems may form in the thyroid system GN.
Availability: GeneNet is available via the Internet at http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/ .

Introduction

The majority of processes in the body reflect a concerted expression of a multitude of genes composing gene networks (GN). The major GN elements are (i) transcription factors, regulating transcription of genes; (ii) cassettes of genes, transcribed in a concerted manner under the action of particular transcription factor; (iii) receptors, providing supracellular communication via interactions with hormones and other signal molecules; and signal transduction pathways providing the intracellular communication. Each GN is formed during the evolution to perform its specific function. Numerous local GN form a hierarchical global GN of the entire organism (Ananko et al., 2002).
To regulate expression of genes adequately, the GN should be capable of (i) rapidly "weeding out" the information that is unnecessary or inessential for the moment; (ii) determining the completeness of the topical information; and (iii) performing adequate actions, including the situations when the necessary information is incomplete. Thus, the GN should possess the capability of functioning in a paradoxical situation of insufficiency of the topical information combined with excess of the overall inflowing information. Moreover, it is important that the size of actual GN exclude principally the possibility of exhaustive search of all the possible decision variants due to their tremendous number.
The nervous system functions in a similar manner. To explain the principles underlying the operation of a body regulated by nervous system, P.K.Anokhin proposed a theory of functional systems (TFS). "TFS is based on the postulate that the backbone factor organizing the functional system of any level is the adaptive result profitable for the organism and the overall system. It is the result that, due to a constant afferent feedback informing the targets on its state, underlies a king of "mobilization" of the central and executive objects into a functional system" (Sudakov, 1984). According to this postulate, all the results are hierarchically ordered while achieving a particular goal. If a motivation dominating at a particular moment requires achieving certain goal and result, the rest results from a hierarchy facilitating implementation of this goal. This is performed in an automatic manner as follows: if achieving of the goal is performed by a functional system, which spends certain resources while its operation resulting in disturbing the equilibrium of certain metabolic constants, the functional systems that maintain these constants at specified levels are automatically switched on. Consequently, all the functional systems form a hierarchy providing attainment of the presently target goal, which becomes at the moment the dominating motivation (Sudakov, 1984). This work considers the organizational and operational principles of real GN in the context of TFS by the example of thyroid system GN.

Materials and Methods

The thyroid system GN described in the GeneNet database was analyzed. This GN is one of the largest and comprises 87 genes, 139 proteins, 81 RNAs, 11 processes, and 410 links between the constituent objects (Suslov, Ignatieva, 2002). It describes the regulation of thyroid hormone syntheses at the intracellular, follicular, hypophysial, and hypothalamic levels

and contains the data on regulation of thyrocyte proliferation and apoptosis. This GN is logically analyzed to interpret its control processes according to TFS.

Results

Five hierarchical levels of the functional systems composing the thyroid system GN may be detected: (i) the level of signal molecules; (ii) the level of local gene networks; (iii) the level of cellular gene networks; (iv) the level of gene networks of the organ; and (v) the level of the overall organismal gene network.

The level of signal molecules. Only one goal*—increase in the signal intensity—is possible at this level. For example, thyroid transcription factor-1 (TTF-1) binds to its site, localized to promoter of it own gene and self-activates. The signal specific of the complex thyrotropin–thyrotropin receptor (TSH–TSHR) is increased via the G-proteins ands adenylate cyclase (Fig. 1,I; Suslov, Ignatieva, 2002).

The level of local gene networks. Signal molecules switch on the functional systems belonging to the second hierarchical level—local gene networks. The goal here is to transduce the signal to the structures specified by genetic memory, and it is attained using the principle of cassette regulation. The following local GNs are described for the thyroid system: TTF-1–dependent synthesis of thyroid hormones (THs); TSH-dependent synthesis of THs; and expression of TSH receptor on the cell membrane.

1) TTF-1–dependent synthesis of THs: the cassette of TTF-1 comprises the genes coding for thyroglobulin (TG), thyroperoxidase (TPO), and sodium–iodine symporter (NIS). Their co-expression underlies the synthesis of iodinated thyroglobulin, forming the complex with asialoglycorpotein receptor of the apical membrane. As a result, the TTF-1 gene expression is inhibited (Fig. 1, II; Suslov, Ignatieva, 2002).

2) TSH-dependent synthesis of THs: the cassette of TSH comprises the same genes (TG, TPO, and NIS) and the gene of H2O2-generating protein. As a result, triiodothyronine (T3) inhibits through its receptor the expression of TSHR gene (Fig. 1, II; Suslov, Ignatieva, 2002).

3) Expression of TSHR on the cell membrane: the cassette of TSHR comprises the gene CREM, one of whose products, ICER, inhibits the expression of TSHR gene (Fig. 1, II; Suslov, Ignatieva, 2002).

The level of cellular gene networks. At the third level—the cellular level, the goal is to commutating the signals between the local GNs. The physical background of this process is formed by (i) organization of the regulatory regions as cassettes of transcription factor binding sites composite elements and (ii) a relay race of the signal molecules. In the first case, switching over is automatic due to different affinities of individual factors for the same site, mutual arrangement of the sites relative to the transcription start, interaction of the factors within a composite element, and differences in concentrations of various factors determined by the history (Ananko et al., 2002). In the second case, the final products remaining untracked by one local GN serve as the signal molecules for the other GN. Examples of the first type are (1) interaction between the TSH- and TTF-1–dependent syntheses of THs, performed via the transcription factor Pax-8, whose gene is activated by TSHR, and (2) regulation of TSHR receptor expression on the membrane through TTF-1, whose binding site is localized to TSHR gene (Suslov, Ignatieva, 2002). Example of the second type is the activation of glutathione peroxidase and superoxide dismutase in response to an increase in the intracellular level of H2O2, diffusing into the cell through the apical membrane and, thus, being a side product of TH synthesis (Fig. 1, III).

Interactions of the local GNs allow the cell functional status to be evaluated integrally (so to say, "from beneath") and save the cell automatically from the self-destruction.

The level of the organ GNs. At this level, the functional state of the cell may be evaluated "from above" according to the levels at which it expresses specialized signal molecules (cytokines, prostaglandins, etc.). Expression of such molecules allows the cells to interact though paracrine and autocrine regulations. Interactions between FAS-L and FAS receptor exemplify the interactions at the level of entire organ. Normally, their expression is balanced by cytokines to prevent an excessive apoptosis. Disbalancing in the cytokine regulation due to appearance of lymphocytes in the thyroid gland during an autoimmune disease leads to a drastic increase in the FAS-L expression. The interaction between FAS-L and FAS receptors triggers apoptosis not only in the thyrocytes actively expressing FAS-L (suicide), but also in the neighboring thyrocytes carrying FAS receptors (fratricide). Cell division aimed to compensate for the apoptosis via inhibition of the endogenous somatostatin is switched in a similar manner (Fig. 1, IV).

The level of the overall organismal GN. The GN of the overall body determines the goal of the highest hierarchical level. The same GN determines the correspondence between the result and the goal. The goal of the GN in question is maintenance of the basic metabolic level; the result, the free T3 blood level. The basic metabolism depends on many characteristics; they are summarized at the level of organs, hypothalamic, and suprahypothalamic levels. Of all these characteristics, GeneNet contains description of the organism's proliferative status, determined according to the levels of growth hormone (GH) and IGF-1. T3 facilitates the release of hypothalamic somatostatin, which, in turn, inhibits the

* We would like to underline once again that speaking of gene networks, we have in mind that the goals and the means for their attaining are formed during the evolution. In the case of a more labile nervous system, the goals are formed by motivations, while the means for their achievement, though also specified evolutionary, are to a considerable degree tuned in the course of training by means of emotions (Vityaev, 1997).

synthesis of TSH and, finally, T3. On the other hand, GH activates the synthesis of IGF-1, activating the syntheses of T3 and somatostatin. Thus, a negative feedback circuit is formed that connects the organism's proliferative status with the level of basic metabolism (Fig. 1, V).

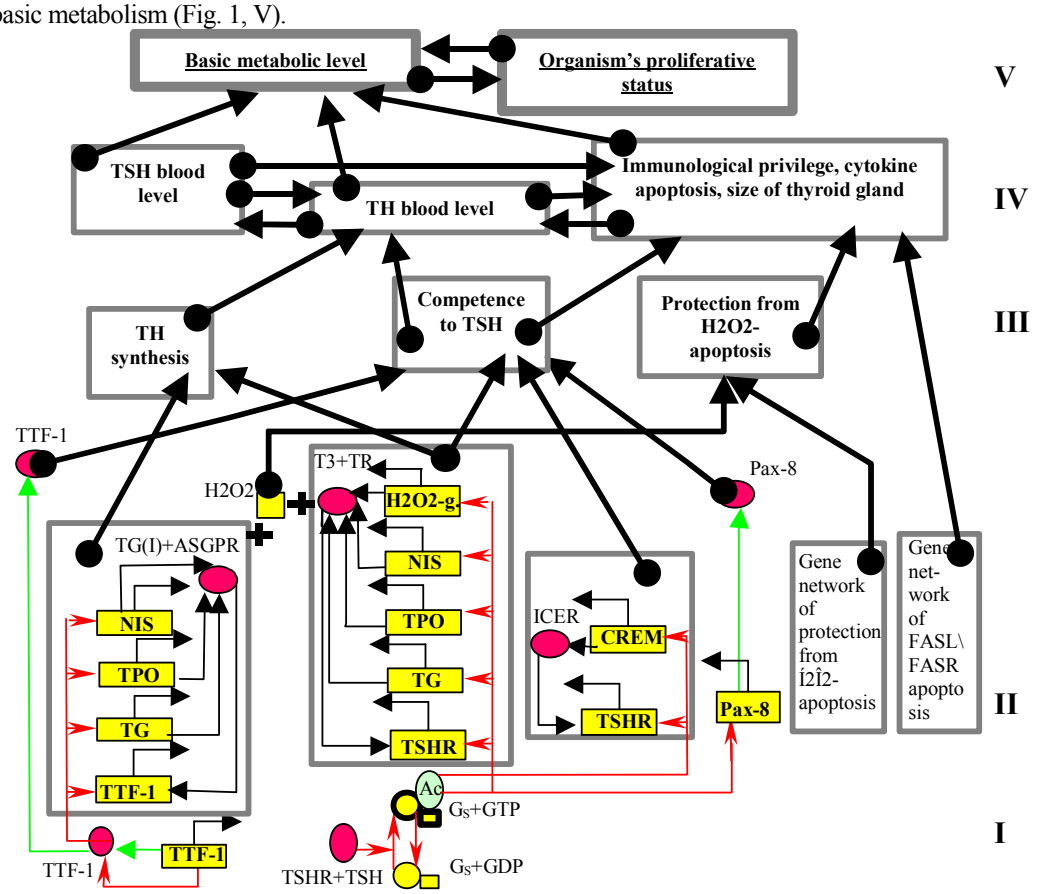


Fig. 1. Hierarchical levels of the functional systems in the gene network of thyroid system.

Discussion

The main merits of the GNs following the TFS principles in their operation are (i) the possibility to tune each local GN independently and (ii) the possibility to rearrange in an adaptive manner the interactions between local GNs at the lower hierarchical level depending on the compliance of the result obtained with the goal specified at the higher hierarchical level. The former stems from the fact that each local GN produced *its own* result and "checks" its correspondence to *its own* goal. The latter results from the interactions of local GNs at the level of metabolic constants (Vityaev, 1997). Let us consider the organization of the functional system within the thyroid GN in the case of deviations in TH blood level (Fig. 2A) and TSH blood level (Fig. 2B).

Deviation in the blood level of thyroid hormones. The TH level may be independently stabilized at hypophysial (level IV), hypophysial–hypothalamic (level IV + V) levels, and via the activation of thyrocyte GN (levels III + II). In the first case, a drop in the TH level results in an increase in the level of hypophysial TSH. TSH stimulates supplementary synthesis of TH. In the second case, a decrease in the TH level causes an elevation in the level of hypothalamic TRH (TSH releasing hormone), thereby stimulating additional synthesis of TSH. TSH stimulates supplementary synthesis of TH. In the third case, a decline in the TH level increases the competence towards TSH (derepression of the gene TSHR). The consequent increase in the number of receptors on the thyrocyte cell surface allows the TH synthesis to be activated even at a low TSH blood level. Having bound to its receptor, TSH activates the TSH-dependent TH synthesis and expression of Pax-8. In turn, Pax-8 activates the TTF-1–dependent TH synthesis. With increase in the blood level of thyroid hormones, their synthesis is inhibited through a negative feedback. Organization of this functional system is given schematically in Fig. 2A.

Deviation in the blood level of thyrotropin. The TSH level may be independently stabilized at hypophysial (level IV) and hypophysial–hypothalamic (level IV + V) levels. In addition, the goal of the highest hierarchical level—the basic metabolism—may be achieved without any additional synthesis of TSH through increasing the competence of thyrocytes towards TSH (levels II + III). In the first case, a decrease in the TSH level leads to a drop in the TH level. The hypophysial negative feedback is weakened, thereby stimulating the TSH synthesis. In the second case, a decline in the TSH level also results in a decrease in the TH level. The hypothalamic negative feedback is weakened, causing an increase in the TRH level,

which stimulates additional TSH synthesis. In the third case, a drop in the TSH level also causes a decrease in the TH level, weakening the negative feedback inhibiting the TSHR synthesis in the thyrocyte. Correspondingly, its competence towards TSH is growing. In turn, the increase in the number of receptors on the thyrocyte surface allows the TH synthesis to be activated even at a low TSH blood level. Organization of this functional system schematized in Fig. 2B.

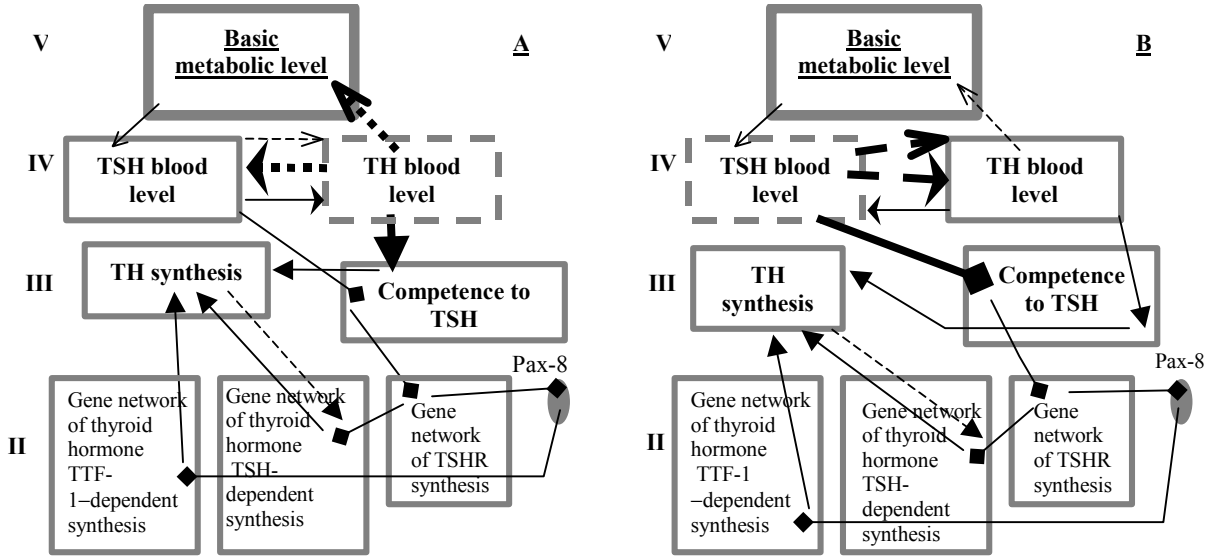


Fig. 2. Organization of the functional system of thyroid GN in the case of deviations in (A) TH and (B) TSH blood levels: broken line, negative effects; solid line, positive effects; arrows with similar ends correspond to the same signal transduction pathway; and bold, initial stimuli that organize the functional system.

Acknowledgements

The work was supported by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90359, 00-04-49229, and 00-04-49255); Russian Ministry of Industry, Science, and Technology (grant № 43.073.1.1.1501); US National Institutes o f Health (grant № 2R01-HG-01539-04A2); and US Department of Energy (Grant № 535228 CFDA 81.049), The authors are grateful to Prof. N.A.Kolchanov for helpful discussions; I.V.Lokhova and L.V.Katokhina, for bibliographical support; and G.B.Chirikova, for translation of the manuscript into English.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002). GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398-401.
2. Sudakov K.V. The General Theory of Functional Systems. M.: Meditsina, 1984.
3. Suslov V.V., Ignatieva E.V. (2002). Molecular genetic mechanisms regulating the thyroid system: description in the TRRD and GeneNet databases. This volume.
4. Vityaev E.E. (1997). Statement of goal as a principle of brain operation. In: Models of Cognitive Processes (Vychislitel'nye sistemy No. 158), Novosibirsk, 9-52.

ORGANIZATION OF THE GENE NETWORK OF APOPTOSIS

**** Stepanenko I.L., Grigor'ev S.A.***

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

Key words: *gene networks, apoptosis*

Resume

Motivation: Death of individual cells is a necessary condition for the maintenance of multicellular organisms. The disruption of normal cell death regulation leads to autoimmune diseases, cancer, and Alzheimer's disease. In the last decade, molecular biologists showed an increasing interest in programmed cell death, which is morphologically classified as apoptosis, and collected large amounts of various experimental data. Computer processing and formalizing can help to systematize and analyze the data collected.
Results: The GeneNet database accumulates information on the structure–function organization of the gene network responsible for apoptosis.
Availability: Apoptosis http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet

Introduction

Apoptosis (programmed cell death) is one of the variations of physiological cell death indicated by recognizable morphological changes (cell shrinkage, membrane blebbing, and DNA fragmentation). Apoptosis is a conservative, highly regulated mechanism for elimination of cells that impair proper functioning of an organism (Hengartner, 2000). Apoptosis is critical for normal development during embryogenesis (Meier et al., 2000), maintenance of tissue homeostasis, and protection from autoreactive T-lymphocytes, virus-infected cells (Krammer, 2000), and tumor cells (Lowe, Lin, 2000). In the last decade, molecular biologists showed an increasing interest in apoptosis. Large amounts of experimental data on programmed cell death necessitate processing and formalization of the data for further computer analysis.

Methods

We used a computer technology employed in the gene network database GeneNet (Ananko et al., 2002). The data visualization program GeneNet viewer displays automatically formalized data on the structure–function organization of the gene network as a complex graph. The nodes of this diagram are regulatory and structural proteins and genes coding for these proteins, and arcs are bonds between them. The information collected by annotating scientific literature is stored in the database in 14 interlinked tables. An SRS-version of the GeneNet database allows a user to extract a list of entities involved in the gene network and a list of all regulatory relations, browse the information about all relationships in the proteins studied, view reactions with proteins, study the role of the proteins in these relations, etc. (Ananko et al., 2002).

Results and Discussion

This database presents main components of programmed cell death (PCD) and shows major ways for activating the PCD mechanism. The gene network of apoptosis is a complex structure that includes a large number of objects linked by various pathways, leading to apoptosis. Specific of the gene network of apoptosis is the prevalence of positive feedbacks, which ensure signal amplification and transduction of the system into a terminal state, i.e., cell death.
Protein–protein interactions and post-translational regulatory mechanisms play an essential role in apoptosis. The genes of main structural components of the gene network are constitutively expressed, and the proteins are present in the cytoplasm in an inactive form. The gene network is activated instantaneously when a certain signal level is reached.
Caspases: key effectors in apoptosis. Caspases, members of the family of cysteine proteases, are synthesized as inactive proenzymes. They are divided into apoptotic initiators and effectors (executioners) basing on their roles in apoptosis. Caspases are activated in cascades by proteolysis with upstream caspases, which contain death domains and are in turn activated by interactions with adaptor proteins. Activated initiator caspases 8, 9, 10, and 12 cleave and activate downstream effector caspases 3, 6, and 7, which, in turn, cleave death substrates and induce apoptosis. Effector caspases initiate either activation of proapoptotic proteins or inactivation of the proteins necessary for the maintenance of the structural integrity of the cell and its survival.
Apoptosis Inhibitors. Activation of the gene network is controlled by apoptosis inhibitors and heat shock proteins (Stepanenko, 2001). IAP and c-IAP-1 binding to c-IAP-2, XIAP, and survivin inhibit caspases 3, 7, and 9. FLIP impedes the interaction between caspase 8 and the Fas-associated death domain adaptor protein. Smac/DIABLO released from the mitochondrion binds to the XIAP inhibitor, thus activating caspases.
The Bcl-2 family of proteins: apoptosis mediators. Activation of caspases is regulated by the Bcl-2 family of proteins that includes survival-promoting proteins (Bcl-2, Bcl-x, Mcl-1), death-promoting members (Bax, Bak, and Bok), and

BH3-only proteins (Bad, Bik, Bid, Bim, and Noxa). BH3-only proteins bind to other Bcl-2 proteins to either inactivate Bcl-2-like proteins or modulate Bax-like proteins. Bcl-2 proteins form pores in the outer mitochondrial membrane, through which cytochrome C is released. Heterodimerization of pro- and antiapoptotic proteins (Bax/Bcl-2) blocks pore formation and apoptosis.

Activation of the apoptosis gene network. In mammalian cells, caspase-dependent apoptosis can proceed via two main pathways: the death receptor pathway through the cytoplasmic membrane and the mitochondrial pathway. However, these pathways converge, thereby activating caspase 3. After that, various subprograms of cell destruction are started. Binding of a ligand with the receptor and adaptor proteins leads to an increase in the local concentration of procaspases and to activation of caspases 8 and 10. Other caspases are activated by upstream caspases, thereby amplifying the proapoptotic signal. Caspase 9 is an initiator in the mitochondrial pathway. Under various stresses, caspase 9 is activated during its binding to the Apaf-1 regulatory proteins and cytochrome C within the complex called apoptosome. An integrator of these two pathways is Bid, whose cleavage by caspase 8 and further translocation to the mitochondrion intensify the release of cytochrome C from the mitochondria. The third compartment involved in apoptosis is endoplasmic reticulum (ER). Various kinds of ER stresses, including Ca^{2+} release, lead to the activation of procaspase 12.

However, apoptosis may proceed without involvement of caspases. During early embryogenesis, apoptosis may result from the release of the apoptosis-inducing factor (AIF) (oxidoreductase) from mitochondria, which obviously acts through an unidentified nuclease. Since apoptosis regulation becomes more and more evolutionarily complicated and the number of caspases and Bcl-2 proteins increases, this pathway for apoptosis activation is phylogenetically the most ancient one. Obviously, the more recent pathway is the ligand–receptor one, which is involved into the immune response. Various ways of apoptosis activation are not isolated from one another: the pathways leading to signal amplification can intersect or combine.

p53-induced apoptosis. p53, a product of a tumor suppressor gene, is an entry of the gene network with a large number of links, regulating various signal-transduction pathways in apoptosis. Let us consider a fragment of the gene network that regulates p53-mediated apoptosis, which combines transcriptionally dependent and independent mechanisms (Fig.).

Normally, the level of p53 in a cell is regulated via the p53-Mdm2 circuit with a negative feedback. p53-mediated apoptosis is activated when the p53 protein level in the cell is increased. The signal for this process can be either a genomic DNA damage or the formation of ROS (reactive oxygen species). The protein is activated as a result of its post-translational modification. The p53 transcription factor induces genes with proapoptotic functions, such as Bax, NOXA, P53AIP, and Fas/Apo-1/CD95, and inhibits transcription of the antiapoptotic gene Bcl-2. If the amount of Bcl-2 protein is sufficient in the cell, this protein normally inhibits potential p53-dependent apoptosis. The Bax protein in the form of active homodimers incorporates into the mitochondrial membranes and forms pores. The translocation of Bax to the mitochondrial membrane results in release of Apaf-1 mitochondrial factors and cytochrome C, which, in turn, activate caspase 9 and induce apoptosis.

The enhancement of Fas gene expression leads to an increase in Fas receptor on the cell surface, and, as a result, the cell becomes competent to FasL-induced apoptosis via cleavage of procaspase 8.

p53 activates the transcription of PIG1–PIG12 genes (p53-induced genes) that are related to the formation of ROS in the cell. The PIG3 gene expression leads to an increase in the production of ROS by the mitochondrion. ROS causes changes in mitochondrial permeability, such as opening of pores in the mitochondrial membrane. As a result, the mitochondrion releases the AIF factor responsible for PCD. Generation of high doses of ROS leads to a decrease in $\Delta\psi$ of the mitochondrial membrane and, hence, apoptosis.

Thus, there are two possible pathways of p53 transcription-dependent apoptosis, which can often be combined. The first, caspase-dependent pathway implies an increase in the p53 expression of the Fas death receptor, induction of the proapoptotic protein Bax, and inhibition of the transcription of the antiapoptotic protein Bcl-2. The second pathway, caspase-independent, implies an increased ROS level, which can result in apoptosis activated by factors of mitochondrial origin. There is also the third pathway that does not depend on the p53 function as a transcription activator. p53 directly activates caspase 8 within an unidentified complex without involvement of the FADD adaptor.

Apoptosis is a complexly regulated process induced by various stress stimuli and exhaustion of survival factors that act via various signal transduction mechanisms. In a gene network, various domains can be simultaneously activated and different signal-transduction pathways can be involved. The cascade principle of signal amplification and integration of signal-transduction pathways underlie the gene network organization. Accumulation of relevant information will improve our understanding of functions of the gene network in different types of cells under various factors leading to apoptosis and reveal the most important nodes in the network, which can be used to regulate the process. Data formalization and systematization in the GeneNet database allow us to perform computer analysis and construct mathematical models.

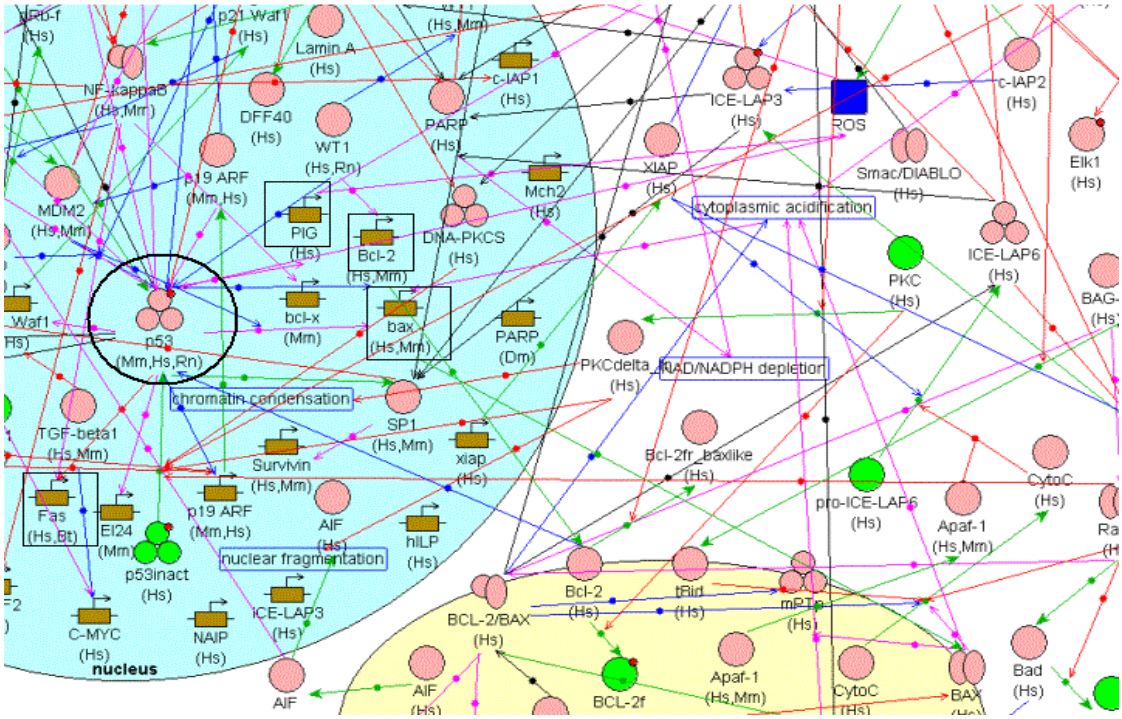


Fig. Fragment of the gene network of apoptosis regulation. The p53 transcription factor induces activation of the Fas death receptor and the Bax–PIG3 mitochondrial pathways to apoptosis.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002) GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398-401.
2. Hengartner M.O. (2000) The biochemistry of apoptosis. Nature. 407:770-776.
3. Krammer P.H. (2000) CD95's deadly mission in the immune system. Nature. 407:789-795.
4. Lowe S.W., Lin A.W. (2000) Apoptosis in cancer. Carcinogenesis. 21:485-95.
5. Meier P., Finch A., Evan G. (2000) Apoptosis in development. Nature. 407:796-801.
6. Stepanenko I.L. (2001) Interference of apoptosis and heat shock response gene networks. Mol. Biol. (Mosk). 35:1063-1071.

GENE NETWORK OF MACROPHAGE ACTIVATION UNDER THE ACTION OF INTERFERON-GAMMA AND LIPOPOLYSACCARIDES

**** Nedosekina E.A., Ananko E.A.***

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

Key words: *gene networks, macrophage activation, database*

Resume

Motivation: As known, macrophages perform different functions in an organism. Disturbance of macrophage functioning is characteristic for many pathological conditions. However, the mechanisms that determine the functioning of this type of cells, as well as the reasons causing pathologies, are still investigated poorly.
Results: Within the frames of the GeneNet system, we have developed a formalized description of the gene network on macrophage activation under the action of lipopolysaccarides (LPS) of bacterial cell wall and the interferon-gamma (IFN-γ). This description contains information about more than 400 components, including 130 proteins, 35 genes, over 200 reactions and regulatory impacts. Visualization of data in the graphical form enables to reveal both the scheme of general structure-functional organization of this gene network and several sub-schemas that represent in much details the signal transduction pathways (e.g., activation of transcription factor NF-κB, Jak-Stat pathway, MAP kinase cascade).
Availability: The gene network on macrophage activation is available via the Internet by the address: http://wwwmgs.bionet.nsc.ru/systems/MGL/GeneNet/

Introduction

Macrophages are important components of the immune system. They participate in the organism's defense from different infections, act in regeneration of damaged tissues, regulate the functioning of other cells, etc. To perform the most of these functions, the macrophages should be activated. In the activated state, the macrophages synthesize numerous proteins and non-proteinaceous substances. Regulation of synthesis of these substances is supported by the gene network on macrophage activation.
The process of macrophage activation is studied rather intensively. Nowadays, the information is available that the gene network on macrophage activation involves hundreds of genes, proteins, and non-proteinaceous substances. Despite the large bulk of experimental information, many details of this biological process are not understood or studied yet.
The goal of the present work was to collect the information from scientific publications about activation of macrophages under the action of two agents, lipopolysaccarides (LPS) from bacterial cell walls and interferon-gamma (IFN-γ) synthesized by T-cells of the immune system and, next, on the basis of this information, to develop the formalized description of the gene network on macrophage activation by using the technology GeneNet (Kolpakov et al., 1998; Kolchanov et al., 2000).

Methods

For the formalized description of the gene network on macrophage activation, we have applied the technology GeneNet (Kolpakov et al., 1998), which was developed for accumulation of information about gene networks and visualization of this information in the graphical mode via the Internet.

Results and Discussion

For developing the formalized description of the structure-functional organization of the gene network on macrophage activation under the action of LPS and IFN-γ, we have used several hundreds of published experimental papers. The current release, dated by April 23, 2002, the information is accumulated on more than 400 different gene network's components, including about 130 proteins, 35 genes, over 200 reactions and regulatory impacts. To simplify the general viewpoint of macrophage activation, some transduction pathways were isolated into separate schemas (e.g., the pathway of activation of the NF-κB transcription factor, Jak-Stat signal transduction pathway, and MAP kinase cascade).
For the full-value functioning, the macrophages should be transformed into the activated state. As the activating agents, different substances could act: lipopolysaccarides (LPS), lipoproteins, hyaluronic acid, double strand RNA, interferons (IFN-α, -β, -γ), interleukines (IL-4, -13, -10), etc. We have chosen only two substances: LPS, as the main component of the external cell membrane of gramm-positive bacteria, and IFN-γ, the cytokine secreted by the T-lymphocytes in response to penetration of infection into the organism. The complete activation of macrophages takes place under simultaneous activation by both these agents (Kovarik et al., 1998).

By analyzing this gene network, it could be noted that the key transcription factors typical for macrophage activation are NF-κB, IRF-1, and Stat-1α (Fig. 1). In comparison to the other transcription factors, these factors enhance transcription of the prevailing amount of genes expressed in a macrophage: some cytokines (*IL-1β, IL-6, IL-12p35, IL-12p40, TNF-α, IP-10, IFN-α, IFN-β*, etc.), enzymes (*COX-2, iNOS*), membrane proteins (*ICAM1*), other transcription factors (*ICSBP, IRF-2*), etc. Besides, some other transcription factors participate in macrophage activation: Pu.1, ICSBP, IRF-2, USF-1, c-Jun, CREB, NF-IL6, AP-1, and CEBPβ (Fig. 1).

For a macrophage's functioning, the soluble cytotoxic molecule, NO, is an important factor (Fig. 2). Reaction of NO synthesis is catalyzed by the enzyme, iNOS. The substrates of this enzyme are L-arginine, molecular oxigen, and NADPH, whereas the co-factors are tetrahydrobiopterin, FAD, and FMN (Marletta, 1993). After macrophage activation, NO synthesis is markedly growing. In the course of this process, enlargement of only iNOS enzyme concentration is insufficient for enhancement of NO gene expression. In order to increase the NO level significantly, it is necessary to increase also the amount of the substrates and co-factors. For example, under the action of IFN-γ (Fig. 2), the enzyme, GTP cyclohydrolase I (GCHI), necessary for the tetrahydrobiopterin synthesis (Werner et al., 1990) is being activated (Fig. 2).

Concentration of the substrate, L-arginine, is dependable upon the transporter, cationic amino acid transporter 2 (CAT2). Its expression is also enhanced when the cell is activated by LPS and IFN-γ. Notably, simultaneous action of these activators is known to decrease the amount of the CAT2 protein against the background of increased mRNA amount. However, the mechanism supporting this phenomenon is still unknown (Kakuda et al., 1999).

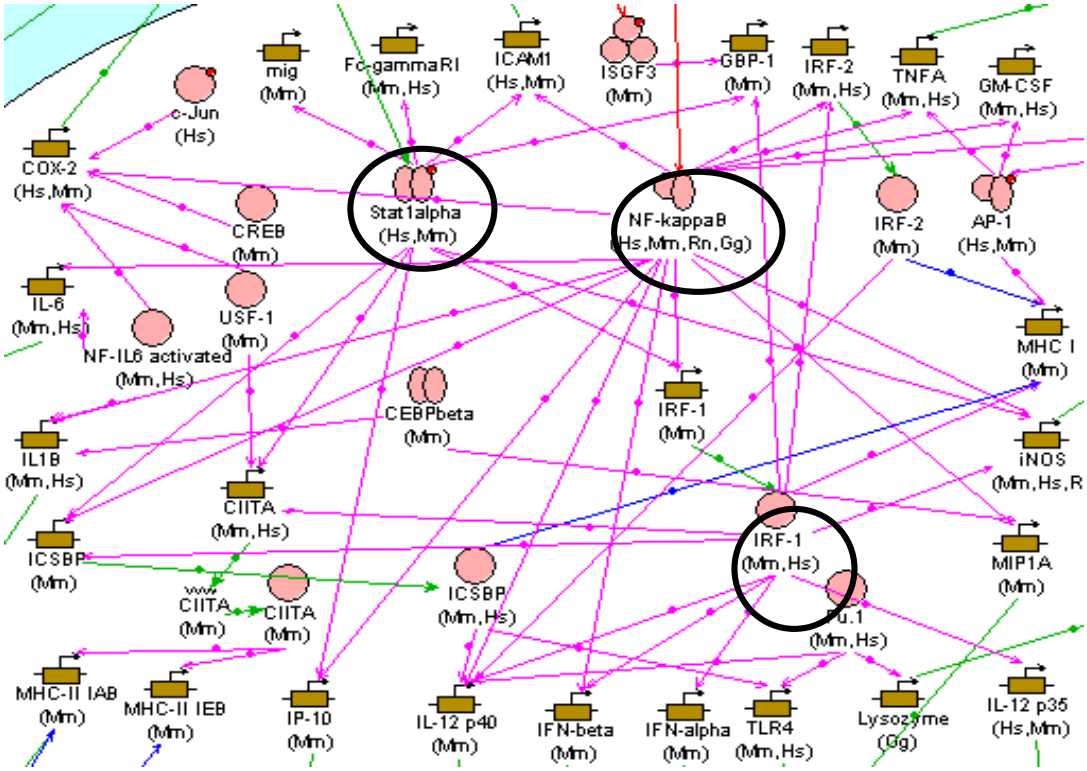


Fig. 1. A fragment of the gene network on macrophage activation: cell nucleus. Filled ovals denote the proteins, in this case, transcription factors (except the transactivator CIITA), filled rectangles, genes; arrows, reactions and regulatory impacts. Key transcription factors are circled.

When analyzing the gene network developed, we have noted the following characteristic features of response of macrophages on the action of LPS and IFN-γ.

1. Action of the LPS and IFN-γ onto a cell starts from different receptors: CD14 for LPS, and IFNR-II for IFN-γ.
2. Both inducers launch the Jak-Stat transduction pathway, which activates the transcription factor STAT1-α. Another transcription factor, NF-κB, is also activated under the action of these both factors. Due to this action, IRF-1 is also activated under the action of both LPS and IFN-γ, because transcription of its gene is activated by both transcription factors, NF-κB and STAT1-α.
3. The sets of genes, transcription of which is stimulated by transcription factors mentioned above are partially intersecting: STAT1-α activates transcription of genes, *iNOS, IRF-1, ICAM1, IP-10, ICSBP, Fc-γRI, mig, GBP-1, CIITA*; whereas NF-κB activates transcription of the following genes: *iNOS, IRF-1,ICAM1, IP-10, ICSBP, IFN-β, IL-6, IL-1β, IL-12p40, MIP-1α, COX-2, IRF-2, TNF-α, GM-CSF*.

4. Joint action of LPS and IFN-γ causes not only strong increase in the iNOS protein concentration (this protein catalyses NO synthesis from L-arginine), but simultaneously decreases CAT2 amount (CAT2 provides arginine transport into the cell), thus, possibly exhausting arginine storages and, in general, decreasing total transcription level in a cell.

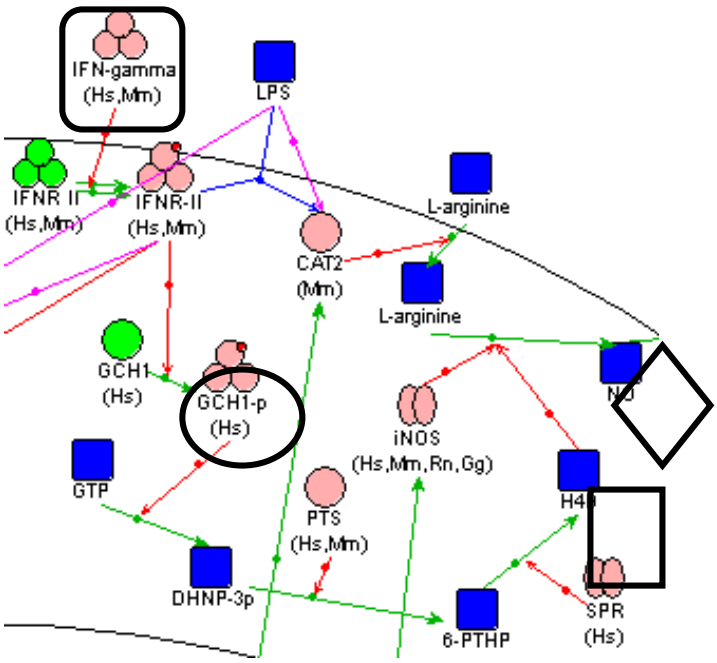


Fig. 2. The fragment of the gene network on macrophage activation. The synthesis of NO and its cofactor, tetrahydrobiopterin. The low molecular weight substances are denoted by filled squares; proteins, by ovals; relations and regulatory impacts, by arrows.

Implementation

Information about the gene network on macrophage activation under the action of LPS and IFN-γ, accumulated in the GeneNet system, could be used for obtaining the integral representation about this process, as well as about the individual role of its particular components. Our work could be useful for the studying of such problems as the influence of particular substances on the macrophage functioning, regulation of myeloid genes expression, the action of mutations on the functioning of the gene network as a whole. On the basis of this information, we have designed the mathematical model of macrophage activation (Nedosekina et al., 2002).

Acknowledgements

The authors are grateful to Lokhova I.V. for bibliographical support.
Work was supported in part by the Russian Foundation for Basic Research (№ 00-04-49229, 01-07-90376, 02-07-90359), Siberian Branch of Russian Academy of Sciences (Integration Projects № 65), Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), National Institutes of Health USA (№ 2 R01-HG-01539-04A2), The Department of Energy USA (№ 535228 CFDA 81.049).

References

1. Kakuda D.K., Sweet M.J., Mac Leod C.L., Hume D.A., Markovich D. (1999) CAT2-mediated L-arginine transport and nitric oxide production in activated macrophages. Biochem J. 340, 549-553.
2. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaia O.A., Ignat'eva E.V., Goriachkovskaia T.N., Stepanenko E.L. (2000) Gene networks. Mol. Biol. (Mosk). 34, 533-544.
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. Kovarik P., Stoiber D., Novy M., Decker T. (1998) Stat1 combines signals derived from IFN-gamma and LPS receptors during macrophage activation. EMBO J. 17, 3660-3668.
5. Marletta M.A. (1993) Nitric oxide synthase structure and mechanism. J Biol Chem. 268, 12231-12234.
6. Nedosekina E.A. (2002) Construction of mathematical model of the gene network on macrophage activation under the action of IFN-γ and LPS. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
7. Werner E.R., Werner-Felmayer G., Fuchs D., Hausen A., Reibnegger G., Yim J.J., Pfleiderer W., Wachter H. (1990) Tetrahydrobiopterin Biosynthetic Activities in Human Macrophages, Fibroblats, THP-1, and T 24 Cells. J. of Biol. Chemistry. 265, 3189-3192.

GENE NETWORK ON CELL CYCLE CONTROL

Turnaev I.I., Podkolodnaya O.A.

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

Key words: *cell cycle, gene network*

Resume[*)]

Motivation: Cell cycle is a key mechanism of cell division, growth, and differentiation. For understanding the mechanisms of governing the cell cycle, it is necessary to standardize available massifs of information on this process, as well as to develop a schema illustrating regulation of the cell cycle in reality.
Results: In this work, a reconstruction of the gene network on cell cycle control represented in the format of the database GeneNet is given. The main functional elements of this gene network are laid down and the general principles of its functioning are being considered.
Availability: http://wwwtest.bionet.nsc.ru/mgs/systems/genenet/

Introduction

Cell cycle is referred to fundamental processes supporting vital activity of organisms. Studying of mechanisms determining the passing of the cell cycle and its effective management are amongst the most pivotal tasks in modern biology. By now, a lot of information is accumulated on cell cycle regulation (Ren et al., 2002, Kohn, 1999, Pines, 1999). The key points of cell cycle regulation are already known, but for understanding the integral schema of the cell cycle management, it is necessary to unify the data available in a single formalized view. In what follows, this approach is useful for developing the model that completely represents the processing of the real mechanism of governing the cell cycle. In this article, the reconstruction of the gene network on cell cycle control (CCC)* is presented together with analysis of information represented in the sections «Cell Cycle G0/G1-S» and «Cell Cycle G2-prophase of mytosis», of the database GeneNet (Kolchanov et al., 2000).

Materials and Methods

Reconstruction of the gene network CCC (G0/G1-S and G2-M(prophase)) was made by applying the GeneNet technology (Kolchanov et al., 2000). The sections of the GeneNet database «Cell Cycle G0/G1-S» and «Cell Cycle G2 – prophase of mitosis» were constructed on the basis of annotating of more than 100 scientific publications. These sections accumulate information about the genes, mRNAs, proteins (including transcription factors), external signals, as well as about regulatory processes. The information used is supplied by the references to the sources of literature, http://wwwtest.bionet.nsc.ru/mgs/systems/genenet/.

Results and Discussion

Reconstruction of the gene network on cell cycle regulation. We have developed the sections «Cell Cycle G0/G1-S» (see the fragment in fig.1) and «Cell Cycle G2- M(prophase)» of the database GeneNet, which contain description of the gene network on regulation of the cell passing from G0/G1 to S phase, together with that from G2 to M(prophase), respectively. In total, these two sections account for 192 objects, 263 relations, 37 genes, 13 mRNAs, 81 proteins, 23 transcription factors (TFs), 7 external signals, 2 non-proteinaceous substances, and 7 processes. The fragment of this gene network is illustrated in Fig.1
Typical features of the gene network on CCC. Analysis of information represented in the sections «Cell Cycle G0/G1-S» and «Cell Cycle G2- M(prophase)» has revealed some typical features of the gene network on CCC and the key mechanisms supporting its functioning. Amongst the most important of them are the following: (1), the central element that provides the concordance in a gene network functioning is the TF, E2F/DP (Fig. 1, 2, 3a, b), which governs by a large gene cassette (Fig. 1, 3a, b); (2), a set of conjugated regulatory contours with positive feedback, each of them making its impact into enhancement of E2F1/DP1 activity, thus, enhancing transcription of the cassette of genes controlled by this TF (Fig. 3c); (3), several contours with negative feedback, which inactivate transcription of the cassette of genes controlled by TF E2F/DP, (elements 3,4,1 in Fig. 4a, 3d, a).

[*)] The following abbreviations are used: CCC, cell cycle control; TF, transcription factor; GF, mitogenic/growth factors.

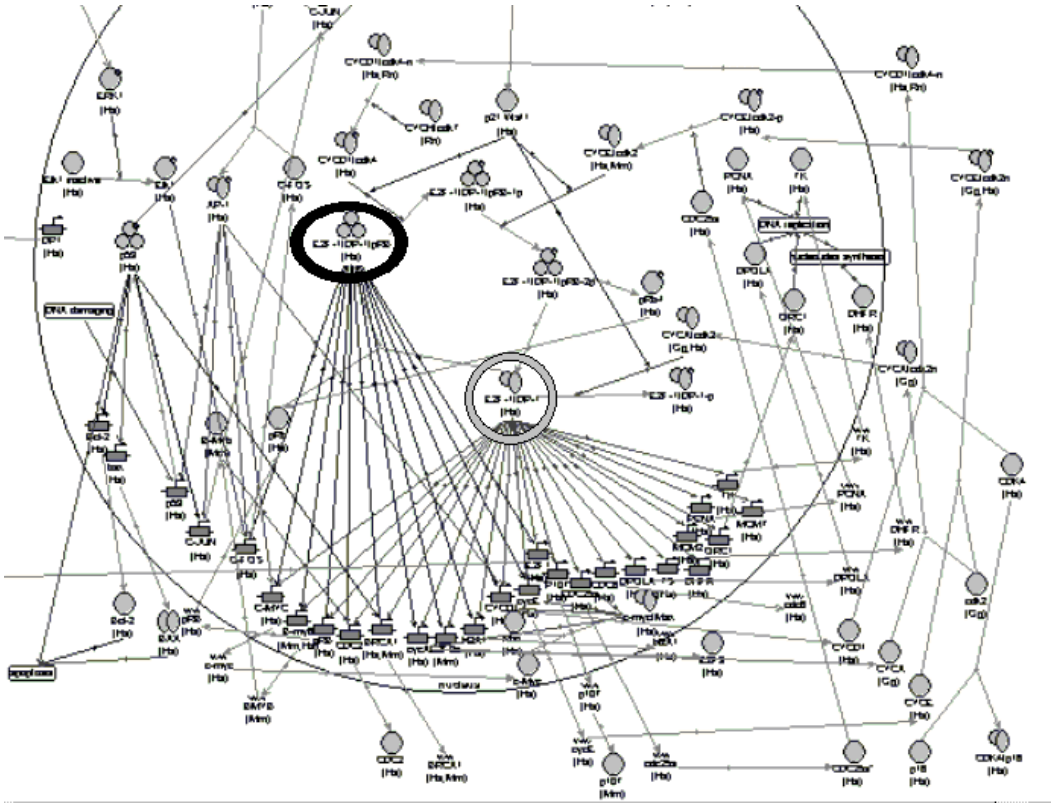


Fig. 1. A fragment of reconstruction of the gene network on cell cycle control, G1-S passage. Black oval, the protein complex suppressing transcription, E2F1/DP1/pRB. Grey circle, the transcription factor, E2F1/DP1, activating transcription.

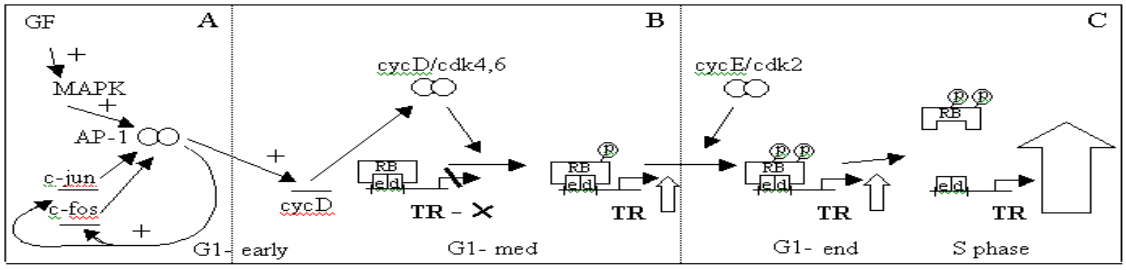


Fig. 2. Stages of the process activating the genes controlled by the TF E2F/Dp: A) Activation by GF, via the MAPK cascade, of the TF AP1(heterodimer jun/fos) and the contour of auto-enhancement of AP1 activation. B) Triggering of the functional regimes of this gene network, first stage, phosphorylation of pRB. By e,d, are denoted subunits of the TF E2F1/Dp1. C) Hyperphosphorylation of pRB causes the final activation of transcription (TR) of genes controlled by E2F.

Triggering of the gene network functioning. The gene network on regulation of the cell cycle control has two regimes of functioning: the stationary one, in a resting cell, and second, the regime aimed at processing of the cell cycle. In the stationary state, the transcription of genes of the cell cycle is blocked by the contour with the negative feedback mediated by the proteins referring to the pocket protein family (G0 and early G1 phases) (Fig. 3a). When a cell passes through a single period of a cell cycle, it is governed by initial activation that is replaced by suppression of transcription of the cassette of genes controlled by the TF E2F/DP (Fig. 3a-b). As a result, the cell proceeds the cell cycle (Fig. 4a,b). The switching between these two regimes, from the stationary one to the execution of the cell cycle (Fig. 3a-b), takes place under the action of extracellular mitogenic/growth factors (GF) (Fig. 2a, b; 3a-b). As a consequence of such switching, the cell passes from G0/G1 to the S phase.

«The element of switching». GFs, by activating the MAP-kinase cascade, induce the TF AP-1 (Fig. 2a). Activation of AP-1 causes the impact on the key link of the gene network on CCC (activation of the cyclin D dependent phosphorylation of the pRB (Fig. 2b)). Activation of the cyclin D and cyclin D dependent phosphorylation of the pRB causes the switching of the pattern of functioning of the gene network on CCC from the stationary state (by blocking of the cell entrance into a cell cycle processing (G0/G1early)) to the state favoring to execution by a cell of the cell cycle program (Fig. 4a, 3a-b). Thus, this link of the gene network (activation of the cyclin D dependent phosphorylation of the pRB) could be denoted as the "switching element" between the regimes of the CCC gene network functioning (elements 1-2 in Fig. 4a, 2b).

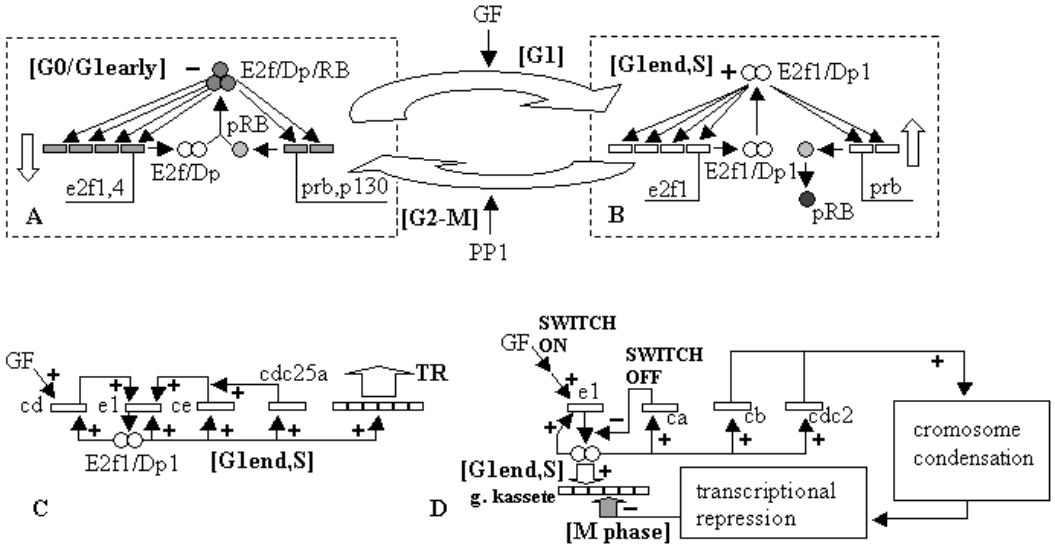


Fig. 3. Regulatory loops with the feedbacks providing the functioning of the gene network: A) The loop with the negative feedback operating in the phases G0/G1early. B) The main contour with the positive feedback that is active in the phases G1end, S. A-B) Direct passing, under the action of extra-cellular growth factors (GFs) (the mechanism is shown in Fig.2) and the reverse trace, the result of dephosphorylation and reactivation of the pRB by the cellular phosphatase PP1. C) The set of conjugated loops with the positive feedbacks. cd, ce are the genes of cyclins D and E; e1, the e2f1 gene. D) Loops with the negative feedbacks operating at the end of the S phase (mediated by the cyclin A) and during mitosis (related to mitotical compactization of the chromatin). ca, cb are the genes of cyclins A and B; e1, the e2f1 gene.

Positive and negative feedbacks. Functioning of the "switching element" switches on the set of conjugated regulatory contours with the positive feedback (element 2 in Fig. 4a, 3c): rapid synthesis of the protein E2F1, enhancement of the TF E2F1/Dp1 activity and, hence enhancement of transcription of the cassette of genes controlled by this TF. Transcription of a part of genes entering this cassette, namely, *cyclin E, e2f1, e2f2, e2f3, cdc25a, pola,* and *pcna,* achieves the maximum at the G1/S, whereas another part of genes, i.e., *h2a1, cyclins A and B,* and *cdc2* is transcribed maximally at S/G2 (Ren et al., 2002). Activation of transcription of a series of genes of this cassette, which encode cyclin A, cyclin B, cdc2, and pRB, after some period of time causes by-stage launching of several regulatory contours with the negative feedback (elements 3,4 and 1 in Fig. 4a, 3a, d). Switching on the negative regulatory impacts leads to damping of gene transcription, and, finally, to returning of the CCC gene network to initial condition (early G1 phase) (element 1 on Fig. 4a, 3a).

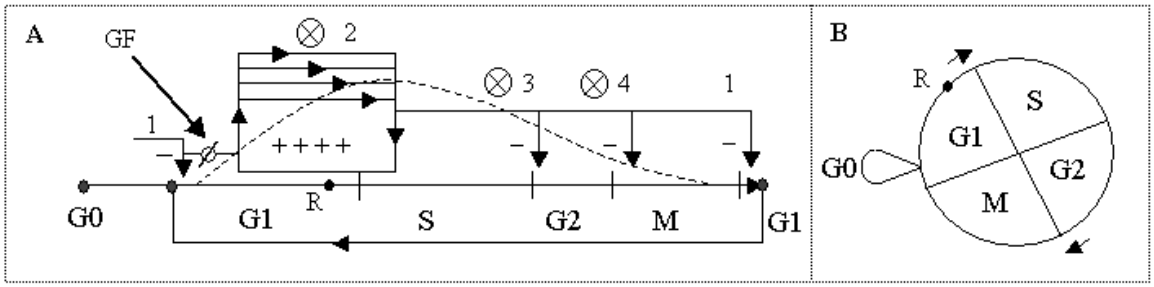


Fig. 4. A) A scheme of operation by the cell cycle. The crossed circle, switching element of the regimes of functioning of the gene network. By numbers are given: the set of conjugated contours with positive feedback (2), contours with negative feedback mediated by RB (1), cyclin A (3), cyclin B, and cdc2 (4). GFs are growth factors; R, R point. Dotted line denotes the plot of alteration of transcription level of the genes controlled by E2F. Cross-wise circle, checkpoints. B) The consequence of phases of the cell cycle. R the point in the late G1 phase, when progression of a cell into S phase becomes irreversible.

Inputs for inner signals. For continuous processing of the cell cycle (cell proliferation), the extra-cellular GFs should be present constantly. The transduction of these signals into the cell nucleus serves as an entrance, which activates cell proliferation (elements 1-2 in Fig. 4a, 2a). The gene network considered has also the entries, activation of which breaks the cell cycle at the checkpoints (checkpoint is a genetic system arresting the passage of a cell to the next stage of the cell cycle until all the processes at the previous stage are not terminated) (Fig. 4a). Under the action on the cell of the antimitogen factors (e.g., TGF-beta) or under DNA damage, activation of inhibitors of cyclin dependent kinases takes place, which suppress activity of kinases and arrest the cell cycle (Fig. 4a).

«Recurrent trigger». The gene network controlling the cell cycle could be conditionally denoted as the "recurrent trigger", because it represents a sort of a device, which is triggered from condition A (element 1 in Fig. 4a) into condition B

(element 2 in Fig. 4a). under the external impact. In its turn, this causes the realization of the working effect (in case of the cell cycle, cell reduplication) with subsequent switching on of the mechanism (elements 3,4,1 in Fig. 4a), which returns the system into initial condition A (element 1 in Fig. 4a). If the system being a recurrent trigger is sensible to the external stimuli only during the period of switching, then the cycle of activation-deactivation will be multiply repeated until the external stimulation terminates. The gene network CCC is sensible to the action of GF only during the period from the beginning of the G1 phase till the R-point (the moment during the mid-late G1, after passing this point, cell progression into the cell cycle is independent from GF, thus, being the result of inner cellular reactions) (Fig. 4a, b).

Conclusion

Thus, in a gene network CCC, the following elements could be considered: 1) MAP-kinase cascade and the contour of multiplication of the AP1, that are destined for the input and enhancement of the external signal (Fig. 2a). 2) The system for enhancement of transcription, via the cassette of genes controlled by the E2F, which consists of a set of conjugated contours with the positive feedback (element 2 in Fig. 4a, 3c). 3) Output working effect, that is, production of two copies from the maternal cell. 4) Recurrent mechanism that is activated after launching the cassette of genes governed by the factor E2F/DP consists of the set of contours with the negative feedback (elements 3,4,1 in Fig. 4a, 3d, a). 5) The mechanism of checkpoints, activation of which leads to switching off regulatory contours with the negative feedback that leads to the interruption of the cell cycle in particular points. This mechanism is launched by extracellular antimitogen factors (for example, TGF-beta) or in the course of damaging some cell structures, for instance, DNA (Fig. 4a). The gene network controlling the cell cycle could be conditionally named as "recurrent trigger". At the next stage of this work, we plan to develop further the sections of the database GeneNet reconstructing the gene network on the CCC, to develop a section of the database on dynamic data on cell cycle regulation. Information presented in these databases will be used for construction of the mathematical model on the basis of the gene network on cell cycle regulation.

Acknowledgements

The authors are grateful to V.A.Likhoshvai for helpful discussions. The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 02-07-90359, 00-04-49229, 00-04-49255), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Project № 65), NIH USA subgrant № 2 R01-HG-01539-04A2.

References

1. Kolchanov N.A., Kolpakov F.A., Podkolodnaya O.A., Ignatieva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000) Gene networks. Russ. J. of Mol. Biol. 34, 533-544 (In Russ.).
2. Ren B., Cam H., Takahashi Y., Volkert T., Terragni J., Young R.A. and Dynlacht B.D. E2F integrates cell cycle progression with DNA repair, replication, and G2 /M checkpoints. (2002) Genes Dev. 16, 245–256.
3. Kohn KW. Molecular interaction map of the mammalian cell cycle control and DNA repair systems. (1999) Mol. Biol. Cell. 10, 2703-2734.
4. Pines J. Four-dimensional control of the cell cycle. (1999) Nat. Cell. Biol. 1, E73-79.

VARIABILITY OF FLOWER DEVELOPMENT GENE NETWORKS IN SEVERAL PLANT SPECIES

**** Aksenovich A.V.***

Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia, e-mail: axenav@bionet.nsc.ru
*Corresponding author

***Key words**: gene networks, flower development, plant species, variability*

Resume

Motivation: During the last decade, the molecular biological data on *gene expression during flower development in various species* are intensively coming, forming a bulk of information requiring *comparison in terms of genetic variability*. The GeneNet computer system provides the tools to realize this goal.
Results: The functions of orthologous homeotic genes in the networks controlling the flower development are amazingly *uniform* in all the plant species yet studied. According to the *patterns of expression,* these genes represent two *groups:* (1) the major group with uniform *expression patterns* and (2) the smaller group displaying dissimilar *time–space expression patterns of orthologous homeotic genes* resulting from differing adaptive strategies of particular plant species.
Availability: The GeneNet module Flower Formation is available at http://wwwmgs.bionet..nsc.ru/systems/mgl/genenet/ as a part of the GeneExpress system.

Introduction

We have earlier described several fragments of the Gene Network of Flower Development in *Arabidopsis* (Omelyanchuk et al., 2001). The data on *gene expression during flower development in various species* are now accumulating rapidly, forming a bulk of information to be *compared in terms of genetic variability*. The GeneNet computer system provides the tools necessary to perform such comparison (Kolchanov et al., 2000). In this work, we found certain specific characteristics of the functions and expression patterns of homeotic genes involved in gene networks underlying flower development in several species, namely:
What remained constant in the gene networks of flower development during over 160 Mya of their evolution and What have changed in the functions and expression patterns of homeotic genes of different species?

Methods

The data for constructing the gene networks mentioned above were taken from the PubMed, MATB, TAIR, and TRRD databases and various publications on mRNA expression patterns of different genes during flower development of plant species and considered in comparison with *Arabidopsis.* For constructing the database, we used the entities “substance”, “gene”, and “protein” and their direct and indirect relations. Only the genes with experimentally identified expression patterns were included into the database developed. A system of filters was used to separate the pieces of information concerning individual species.

Results and Discussion

The *functions* of orthologous homeotic genes in flower development networks are amazingly *uniform* in all the studied 53 plant species belonging to different phylogenetic taxa from the orders Coniferales to Poales.
According to their *patterns of expression,* these genes represent *two groups:*
The major group with uniform *expression patterns* and
The smaller group displaying dissimilar *time–space expression patterns of orthologous homeotic genes* resulting from differing adaptive strategies of particular plant species.
The group with uniform *expression patterns of orthologous homeotic genes* represents the families of orthologs of the following genes:
Gene LFY of Am:FLO (*Antirrhinum majus* L.), Eu:ELF1 (*Eucalyptus globules*), Ps:UNI (*Pisum sativum* L.), ImpFLO (*Impatiens balsamina*), etc.;
Gene AP1 of Am:SQUA (*Antirrhinum majus* L.), Sa:SAAP1 (*Sinapis alba*), etc.;
Genes AP3 and PI of Le:TM6 (*Lycopersicum esculentum*), Am:DEF (*Antirrhinum majus* L.), pMADS1 (*Petunia hybrida*) , NTDEF (*Nicotiana tabaccum*), etc.; and
Gene AG of Pm:SAG1 (*Picea mariana*), Le:TAG1 (*Lycopersicum esculentum*), Ra:RAP1(*Rumex acetosa*), Os:RAG (*Oryza sativa*), etc.
The group displaying differing *time–space expression patterns of orthologous homeotic genes* consists of six classes separated with reference to the factors underlying the alterations in expression patterns of orthologous homeotic genes.

1) The first class of factors change mainly the *time of expression of the flower development gene networks reflecting the changes in the corresponding signals that control* the initiation of flowering, i.e. switch on these gene networks. These factors are listed in Table 1.

Table 1. Changes in the signals that initiate flowering in different plant species.

Genes	Plant species	Type of expression alterations
AP1 orthologs:		
LtMADS1, LtMADS2	Lolium temulentum	Expression begins after 30 h of LD induction in a specialized vegetative cone
AP3,PI orthologs:		Expression begins earlier in a specialized vegetative cone
OSMADS2, OSMADS4	Oryza sativa	
Silky 1 ,ZMM16,		– " –
ZMM18, ZMM29.	Zea mays	
LFY orthologs:		– " –
BM8		Expression begins early in panicle, not in florets
RFL	Hordeum vulgare	
	Oryza sativa	Expression begins relative late (about 12 day after 30 h of LD induction) in a specialized vegetative cone
Lt LFY		
	Lolium temulentum	Induction of expression effects by the exposure to 8 short days (not by a long day as in *Arabidopsis*)
ImpFLO	Impatiens balsamina	Expression is induced only at the age of over 15 years old
		Mutations in the NEEDLY promoter result in the response to photoperiod: it is induced by a short day, whereas the LFY promoter of *A. thaliana,* by a long day
PTLF	Popolus trichocarpa	
NEEDLY	Pinus radiata	

2) Some species have genes of higher hierarchical level that change the site of expression of flower homeotic genes.
a) In *Petunia hybrida, Nicotiana tabaccum,* and *Lolium temulentum,* the expression of cassette activators of flower development gene networks—the orthologous genes ALF, NFL 1, and Lt LFY (respectively)—is initiated in cooperation with the genes determining the type of inflorescence. For example, the genes ALF and EXP of *Petunia hybrida* initiate cymose inflorescences, unlike *Arabidopsis*, which has racemose inflorescences and only one corresponding gene At:LFY. ALF and EXP function in two distinct processes. EXP is required for initiating differentiation of inflorescence meristems into two new meristems types, but has no effect on the particular properties of these meristems. On the other hand, ALF determines the floral type of one of these two meristems, being unable to initiate differentiation of the meristem.
b) In the case of *Lycopersicon esculentum,* vegetative and reproductive phases alternate regularly during sympodial growth. The inflorescences in wild 'indeterminate' plants are separated by three vegetative nodes. Fewer nodes are developed in 'determinate' plants homozygous for the recessive allele of the SP gene until the shoot is terminated by two inflorescences. The floral induction requires two genes: SP gene to regulate the alternation between vegetative and reproductive cycles in sympodial meristems and the FA to determine the floral identity.
3) Many species contain two or more paralogs of flower homeotic genes differing both in time and site of their expression (Table 2).

Table 2. Paralogs of flower homeotic genes in several plant species.

Genes	Plant species	Type of expression alterations
LFY orthologs:		As a result of sex dimorphism, NEEDLY is expressed only in female cones;
NEEDLY, PRFLL	Pinus radiata	PRFLL, only in male cones
		Different site–time expression pattern of paralogs
vcLFY1,vcLFY2	Violet cress	– " –
AP1 orthologs:		
EAP1 , EAP2	Eucaliptus g.	– " –
LtMADS1,	Lolium temulentum	– " –
LtMADS2	*Brassica oler*.	
BoAP1-A,B		– " –
	Silene latifolia	
SLM4, SLM5		– " –
AP3,PI orthologs:		– " –
	Rumex acetosa	– " –
RAD1, Ra:RAD2 DEF1,DEF2	Gerbera hybrida	
STDEFpD12,pD13	Solanum tub.	– " –
ZMM16,		
ZMM18, ZMM29	Zea mays	– " –
OSMADS2, OSMADS		– " –
PLE and FAR	Oryza sativa	
AG orthologs: MASAKO	Antirrhinum m.	– " –
C1,D1		– " –
CAG 1,CAG 2 pMADS3, FBP6	Rosa rugosa	– " –
SLM1, SLM 5 GAGA1,GAGA2	Cucumis sativa	– " –
PTAG1, PTAG2	Petunia hyibrida	– " –
ZAG1, ZMM2	Silene latifolia	– " –
	Gerbera hybr.; Populus trich.; *Zea mays* L.	

4) In the case of an apomictic line of *Hieracium piloselloides,* HPDEF, and ortholog of AP3, does not expressed in the specialized zone of chalazal region, resulting in formation of aposporial embryo sac. The partenocarpy of partenocarpic varieties of *Malus domestica* is a result of specific mutations—insertions of retrotransposon into introns 4 and 6 of the gene MdPI (an orthologs of PI), leading to impaired translation of this gene.
5) Orthologs of AP3 and PI **(**SLM2 and SLM3 in Silene latifolia; PTD in Populus trichocarpa) display different expression patterns in male and female flowers.
6) Only one of the species analyzed—*Medicago sativa*—carries the gene NMH7, which exhibits a pronounced divergence from other orthologs of AP3 and appeared to perform not only the function B in a very changed manner, but also a new function in root nodulation.

Conclusions

The gene network of flower development, including so far the data on 53 plant species belonging to various phylogenetic taxa from Coniferales to Poales, is highly conservative in the functions and patterns of expression of flower homeotic genes. Rare examples of variability in the patterns of expression of flower orthologous homeotic genes are rather due to higher-level regulatory mechanisms that switch on the gene network of flower development and to paralogy of these homeotic genes. The variability of orthologous homeotic genes involved in the gene networks of flower development is a result of variability in their nucleotide sequences (orthologous sequences display a 99–53% similarity). The evolution of homeotic regulatory genes as a result of *nonsynonymous nucleotide substitutions* correlated closely with the phenotypic, adaptive evolution; this solves the key paradox of evolutionary genetics, that is, the disparity between rates of morphological and structural gene evolution (Barrier et al., 2001).

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 02-07-90359, 00-04-49229, and 00-04-49255); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project № 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). Authors are grateful to Prof. N.A.Kolchanov for constant attention and recommendations.

References

1. Barrier M., Robichaux, R.H., Purugganan M.D. (2001) Accelerated regulatory gene evolution in an adaptive radiation. Proc. Natl Acad. Sci. USA. 98(18):10208-10213.
2. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko E.L. (2000). Gene networks. Mol. Biol. (Mosk.). 34(4):533-544.
3. Omelyanchuk N.A., Aksenovich A.V., Stepanenko I.L. (2001). Gene network of flower formation in *Arabidopsis:* its description in the GeneNet system. In: Proc. Intern. Conf. on Genetic Collections, Isogenic and Alloplasmic Lines, Novosibirsk: IC&G SB RAS. 271-275.

NEGATIVE REGULATION OF PLANT PHOTOMORPHOGENESIS

Smirnova O.G., Ibragimova S.S., Shavrukov Yu.N., * Stepanenko I.L.

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

Key words: *gene networks, photomorphogenesis, photoreception, signal transduction*

Resume

Light is one of the most important regulators of plant development. Plants react to quantitative and qualitative light characteristics owing to the system of photoreceptors and a branched network for light signal transmission. The elements of this network, their relationships, and functions are presented in the gene network regulating photomorphogenic development of plants. It is shown that along with positive regulation, negative regulation plays an important role in photomorphogenesis. Several levels of negative regulation are distinguished, which ensure initiation of a photomorphogenic response, limitation of light activation, and optimization of seedling growth under varying light conditions.
Availability: Photomorphogenesis http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet

Introduction

Plant development includes several stages, namely, embryogenesis, vegetative development (formation of stem and leaves), generative development (formation of floral meristems and flowers), and senescence. All these stages, starting from germination, are affected by light (Casal, 2002) as well as functioning of one third of the plant genome, in which three fifth of genes are activated and two fifth are repressed by light (Ma et al., 2001).
Photomorphogenesis is a program of seedling development, in which heterotrophic growth at the expense of internal seed reserves is switched to auxotrophic nutrition of plants with photosynthesis products. Seedlings germinated in darkness are of yellow color and have long hypocotyls, poorly developed cotyledons, and no chloroplasts. After germination, light induces de-etiolation of the seedlings. Studies of pathways of signal transduction from photoreceptors to light-regulated genes are of great interest (Quail, 2002).

Methods

The GeneNet database technology used in this study includes a database and the GeneNet viewer data visualization program (Ananko et al., 2002). The GeneNet Input interface specially designed for this database allows an automatic direct translation of the input information into the GeneNet database format. Formalized data on the structure–function organization of the gene network is displayed as a diagram. The gene network objects and links between them are provided with references to published scientific data and the databases EMBL, SWISS-PROT, PubMed, and TRRD.

Results

The gene network of light-regulated plant development consists of several local networks, which include key genes and regulatory molecules involved in light signal transduction.
The local gene networks regulate functioning of a repression complex in darkness, deactivation of this complex in light, and formation of a photomorphogenic response in the form of inhibition of hypocotyl growth, development of cotyledons and chloroplasts, and anthocyanin accumulation. Local gene networks become available when the corresponding filters are used (Filter: Set filter by inducer/repressor).
COP1 repression complex in darkness. In an etiolated seedling, function of light-activated genes is repressed by a complex with COP1 protein as the main component (Fig. 1).
On the one hand, functioning of COP1 in darkness represses functions of nuclear (cab, rbcS, chs, cip4, cip7) and chloroplast (rbcL) genes, resulting in the growth of an elongated etiolated seedling. On the other hand, the activation of phyA gene transcription by COP1 leads to the high level of inactive photoreceptors in the cytoplasm. The basal level of COP1 activity in darkness is provided by the cytoplasmic protein FIN219 and the IMPalpha1b importin. In the nucleus, COP1 affects ubiquitination and degradation of the transcription factor HY5, which plays an important role in the activation of gene transcription in light.
Inhibition of COP1 activity in light. Three pathways of COP1 deactivation can be distinguished (Fig. 2). The quickest response is due to the direct interaction of photoreceptor molecules with COP1. As a result, the transcription of some light-activated genes is derepressed, including cip4 and cip7, and the next pathway, involving interaction of COP1 with CIP4, CIP7, and SPA1 nuclear proteins, is activated. At the final stage, COP1, which is localized in the nucleus in darkness, is transported from the nucleus to the cytoplasm, where it binds to the CIP1 and CIP8 proteins.

Light activates positive and negative regulators of photomorphogenesis. Many proteins regulating photomorphogenesis known at present are positive regulators. Their expression is regulated by light both positively (HY5, HFR1/RSF1, CIP4, and CIP7 proteins) and negatively (PIF3, FIN219, PRA2, and FHY1/PAT3 proteins).

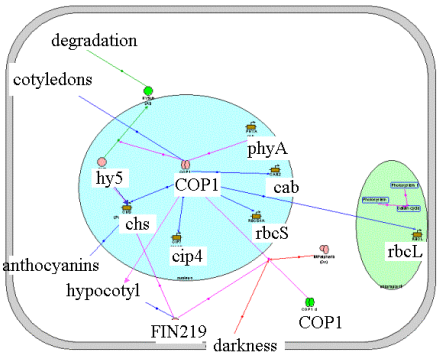


Fig. 1. Gene network of plant photomorphogenesis regulation ("COP1_darkness" filter). The COP1 repression complex in darkness.

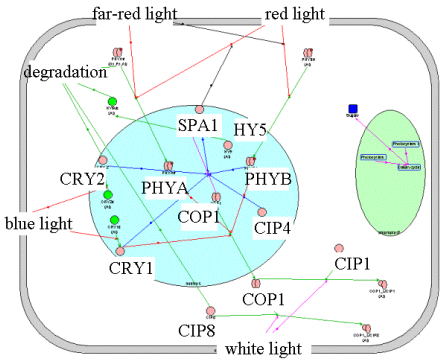


Fig. 2. Gene network of plant photomorphogenesis regulation ("COP1_darkness" filter). Repression of COP1 activity in light.

The level of PHYA и CRY2 photoreceptors is regulated negatively by light. In high-intensity red and blue light, the PHYA and CRY2 photoreceptors degrade and the level of phyA gene transcription decreases. Photostable receptors PHYB and CRY1 contained in small amounts in the seedling start playing an active role. Thus, high-intensity light signals are transmitted via the PHYA and CRY2 photoreceptors. If the light intensity increases, the signal is transmitted by PHYB and CRY1 photoreceptors. Because of negative light regulation of the level of PHYA and CRY2 photoreceptors and several regulatory proteins, the seedling has a limited perception of high-intensity light signals, thus allowing the metabolic pathways to function at an optimal speed.

The transcription factors HY5 and PIF3 are key positive regulators of hypocotyl growth, anthocyanin accumulation, and chloroplast development. Phytochromes initiate two signal-transduction pathways. The PIF3 factor activated by direct interaction with the PHYB photoreceptor is not affected by phosphorylation. The activity of pif3 gene transcription decreases in light. A lengthier signal-transduction pathway is involved into the activation of HY5 protein (Fig. 3). The repression of COP1 in light results in a decrease in HY5 degradation. Phosphorylation affects the HY5 activity. A reduction in protein kinase CK2 activity leads to an increase in the HY5 active nonphosphorylated protein pool. The amount of active HY5 increases in light due to an increase in hy5 gene transcription.

Along with positive regulators, light activates some negative regulators that decrease the intensity of the seedling response to the light signal. SUB1, a light-activated negative regulator, limits the activity of the transcription factor HY5.

Inhibition of hypocotyl growth. Hypocotyl is a part of the stem between the root and cotyledon leaves. Hypocotyl growth in light is inhibited as a result of the balance between positively and negatively regulated cell processes (Fig. 4). Hypocotyl grows intensely in darkness because of involvement of the COP1 repressor and PRA2 and ATHB-2 proteins.

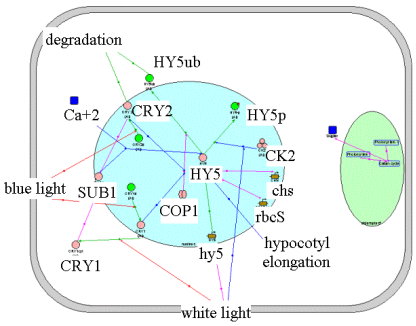


Fig. 3. Gene network of plant photomorphogenesis regulation ("HY5" filter). Activation of the HY5 transcription factor.

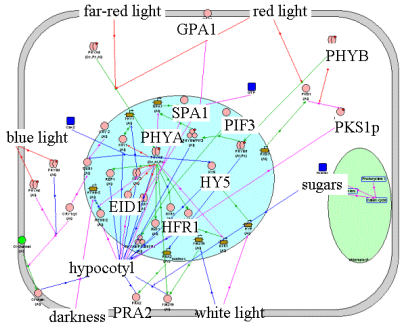


Fig. 4. Gene network of plant photomorphogenesis regulation ("hypocotyl" filter). Inhibition of hypocotyl growth.

Specific mechanisms in hypocotyl that reduce the number of DNA endoreduplication cycles and cell divisions and decrease cell elongation remain unknown. It is shown that in addition to positive regulators inhibiting hypocotyl elongation, some proteins such as SPA1, EID1, PKS1, and SUB1, are active in light that have a negative effect on the decrease in hypocotyl length. FHY1, a positive component of the PHYA signal transduction pathway, can be either a positive or a negative regulator of the PHYB signal pathway of hypocotyl growth regulation.

Anthocyanin accumulation in the cell. Similar limitation of light activation is observed upon regulation of anthocyanin accumulation (Fig. 5). CHS is a key enzyme in anthocyanin synthesis. Along with positive transcription factors CPRF2, CPRF4, CPRF1/CPRF4, and MYB1, light induces accumulation of the CPRF1 factor, which is a negative regulator in the chs and cprf1 gene transcription. Anionic channels play an important role in anthocyanin accumulation. In this case, light activation is also limited: the phytochrome A photoreceptor is a positive regulator and phytochrome B is a negative regulator of activation of anionic channels by cryptochrome 1.

Formation of mature chloroplasts. RBCS and CAB genes involved in the Calvin cycle and the formation of photosynthetic machinery are negatively regulated by protein kinase CK2. CK2 is inhibited in response to light, which leads to activation of RBCS and CAB genes (Fig. 6). CCA1, a cab transcription factor is regulated in response to light by the negative feedback principle.

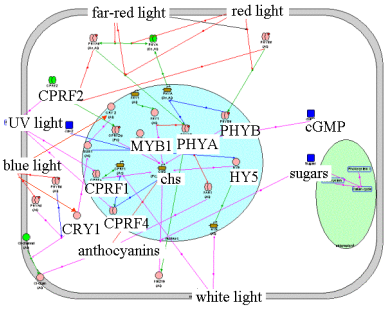


Fig. 5. Gene network of plant photomorphogenesis regulation ("anthocyanin" filter). Anthocyanin accumulation in the cell.

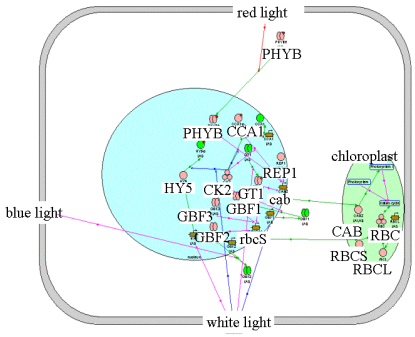


Fig. 6. Gene network of plant photomorphogenesis regulation ("chloroplast" filter). Formation of mature chloroplasts.

Thus, light can activate both positive and negative regulators, whose joint action is aimed at maintaining homeostasis during photomorphogenic development of the seedling.

Discussion

COP1, a negative photomorphogenesis regulator, inhibits activation of light-regulated genes. A similar mechanism of derepression is found at early stages of seed germination. The gene activity in plant seeds is blocked by abscisic acid (ABA), a phytohormon. Water removes a repressive effect of ABA. Thus, the mechanism of blocking the gene system and derepression of this mechanism are observed not only during photomorphogenesis, but also during seed germination. A similar mechanism is found at other stages of plant development, for example, during flower formation. The repression removal that provides a quick response of the system to environmental factors is a key moment in the gene network operation during plant morphogenesis.

Analysis of the gene network of photomorphogenic development suggests that a seedling has a complex system of regulation of photoreceptor and protein activities, which allows optimal perception of light signal. Negative links play a key role in photomorphogenesis. Several levels of negative regulation can be distinguished in this gene network.

The first level regulates function of the pleiotropic gene cop that represses light-activated genes in darkness.

At the second level, the activity of the repression complex is inhibited in light, resulting in changes in the activities of hundreds of genes.

The third level implies limitation of the seedling response to light. This is explained by the effect of local gene networks, leading to an increase in the level of photoreceptors and positive regulators and to light activation of negative regulators.

The forth level regulates transcriptional factors that affect negatively the transcription level of their genes (CCA1, LHY, ATHB-2, and CPRF1) via regulatory circuits with a negative feedback. In the photomorphogenic response, various cell processes are regulated by light both positively and negatively, thus providing an optimal response of the seedling to varying light conditions.

The GeneNet database stores data on gene networks regulating plant development. At present, the database encompasses gene networks regulating degradation of reserve substances during seed germination, seedling de-etiolation in light, nodulation in legumes, flower formation, and biosynthesis of reserve substances during seed maturing. We plan to describe all stages of plant development and construct a mathematical model of plant morphogenesis.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002) GeneNet: a database on structure and functional organization of gene networks. Nucl. Acids Res. 30:398-401.
2. Casal J.J. (2002) Environmental cues affecting development. Curr. Opin. Plant. Biol. 5: 37-42.
3. Ma L., Li J., Qu L., Hager J., Chen Z., Zhao H., Deng X.W. (2001) Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways. Plant Cell. 13: 2589-607.
4. Quail P.H. (2002) Phytochrome photosensory signaling networks. Nat. Rev. Mol. Cell. Biol. 3: 85-93.

A HYBRID NETWORK OF NITROGEN-FIXING NODULES: INTERGENOMIC INTERACTIONS OF BACTERIA AND HOST PLANT

Ibragimova S.S., Smirnova O.G., Shavrukov Yu.N., * Stepanenko I.L.

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

Key words: *gene networks, nodulation, plant–microbe interaction*

Resume

Motivation: In symbiosis with soil bacteria of the genus *Rhizobium*, legumes have a unique ability to fix nitrogen from the air. The GeneNet database accumulates various data on functioning of gene networks and provides reliable representations of hybrid gene networks.
Results: Using the GeneNet technology, the authors described early stages of functioning of a hybrid gene network that controls the symbiotic interaction of two organisms, a host plant and a bacterium. By accumulating the relevant data, the gene network of nodulation regulation will be reconstructed, and the molecular genetic mechanisms underlying the ability of legumes to form nitrogen-fixing root nodules will be detected. This work is of considerable importance for plant agronomy and transgenesis.
Availability: Nodulation http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet

Introduction

All the processes occurring in an organism are due to the coordinate expression of various gene groups forming gene networks (Kochanov et al., 2000). The information on the structure–function organization of gene networks is stored in the GeneNet database (Ananko et al., 2002). The database includes gene networks regulating lipid metabolism, endocrine regulation, cell differentiation, morphogenesis of tissues and organs, and response to external stimuli.
Gene networks function at the cellular level as well as at the level of the entire body. The symbiotic interactions of a host plant with microorganisms lead to the formation of the so-called hybrid gene networks at the next, superorganic level. In a symbiotic gene network, genes of two organisms are coordinately expressed and signal transduction pathways and metabolic pathways are combined.
Legumes, which have attracted much attention, have a unique ability to fix nitrogen in air in symbiosis with soil bacteria of the genus Rhizobium (Schultze, Kondorosi, 1998). In the symbiotic interaction, the following stages are distinguished: (1) preinfection, which includes the induction of genes of bacterial virulence, deformation of root hair, and formation of the nodule meristem; (2) infection of plants and formation of nodules, including the formation of an infectious filament, endocytosis of the bacteria into the host plant cell, and differentiation of bacteroids; and (3) functioning of nodules, i.e., nitrogen fixation.
We are considering the first stage of the formation of the hybrid gene network.

Methods

The GeneNet database technology includes a database and the GeneNet viewer data visualization program (Ananko et al., 2002). The GeneNet Input interface specially designed for this database allows an automatic direct translation of the input information into the GeneNet database format. Formalized data on the structure–function organization of the gene network is displayed as a complex diagram. The gene network objects and links between them are provided with references to published scientific data and the databases EMBL, SWISS-PROT, PubMed, and TRRD.

Results

Nod factor synthesis. Nod factor, a signal molecule of the bacteria, induces important physiological processes and gene expression that are necessary for activation of the nodulation gene network. In response to the colonization of the rhizosphere by bacteria, plants secrete flavonoids, specific root exudates. Flavonoids activate virulence genes, namely, *nod* genes in *Rhizobium* (Fig.). At the first stage of the interaction, the NodD gene transcription in bacteria is enhanced, and the product induces a cassette of other *nod* genes, namely, NodA, NodB, and NodC, that are involved in the formation of the basal structure of Nod factor, *Rhizobium* lipochitooligosaccharide signal molecule. The species specificity of Nod factor is controlled by the bacterial NodE gene, whose host's specificity is determined by the NodH gene.

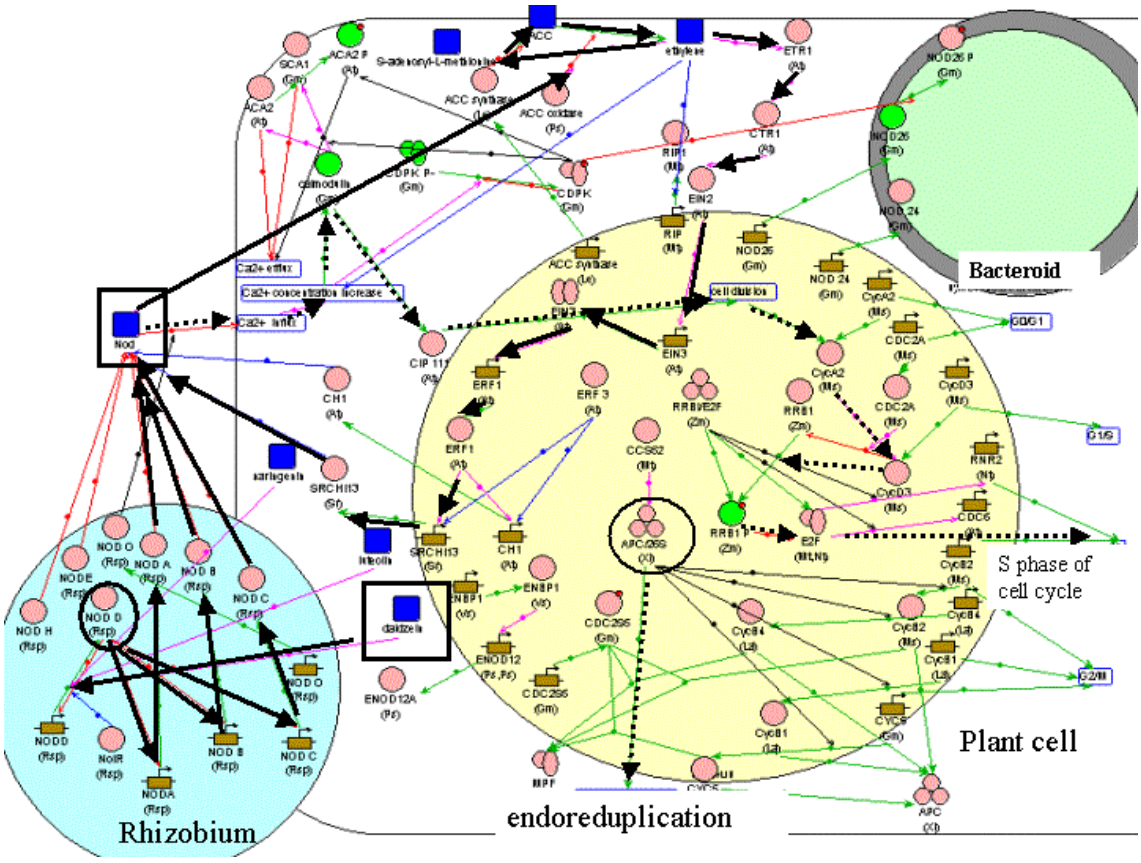


Fig. Hybrid gene network regulating early stages of nodulation in legumes. The arrows show the regulation pathway of Nod factor synthesis, cell cycle reactivation by Nod factor, and the arrest in the G2 phase leading to cell endoreduplication.

Concentration of Nod factor in the rhizosphere is controlled via a regulatory circuit with a negative feedback mediated by the phytohormone ethylene. Nod factor activates the ethylene synthesis pathway. Synthesis of ethylene in the cell is increased via a regulatory control with a positive feedback. The formed ethylene itself is the activator of ACC synthase, involved in the conversion of ethylene precursor S-adenosyl-*L*-methionine into ACC (1-aminocyclopropane-1-carboxylic acid). The ethylene signal is transduced via the ethylene receptor ETR1 and proteins CTR1 and EIN2, localized to the cytoplasmic membrane, and activates the transcription of the gene EIN3. The transcription factor EIN3, in turn, activates the gene ERF1. The key event of the ethylene signal transduction pathway is transcription activation of the major plant chitinases by transcription factor ERF1. The chitinases hydrolyze Nod factor, providing the control of the plant over nodule formation.

The bacterial Nod factor induces the expression of early nodulation genes (ENOD2, ENOD5, ENOD12, ENOD40, etc.) and cell cycle genes in the host plant.

Regulation of Ca2+ accumulation in cells. During first three minutes, the action of the Nod factor results in an increase in Ca2+ concentration around the root hair and in the cytosol followed by regular peaks in the vicinity of the plant cell nucleus.

The Ca2+ signal transduction pathway is regulated by both symbiotic partners. Ca2+ influx into a plant cell is controlled by the bacterial protein NodO, which forms ion channels on the surface of the cell membrane and regulates the Ca2+ influx by blocking the ion channels. Ca2+-dependent ATPases SCA1 and ASA2 regulate the Ca2+ ion concentration in the cell with the help of calcium pumps.

The pumps are activated by calmodulin, the main Ca2+ receptor. The pump activity is regulated by the calcium-dependent protein kinase (CDPK). CDPK phosphorylation inhibits pump activity and binding to calmodulin.

An increase in intracellular Ca2+ concentration due to the Nod factor activity and the efflux of Cl– and K+ ions leads to membrane depolarization, reorganization of the cell cytoskeleton, and cell cycle activation. In one complex with calmodulin, calcium binding to other proteins is involved in the signal transduction pathway of activation of cell cycle genes, particularly, CIP 111.

Cell cycle reactivation. Inner cortex cells blocked in the G0/G1 phase are reactivated by the Nod factor to enter the cell cycle. The transcription factor E2F in complex with the retinoblastoma protein RRB1 is unable to activate S-specific gene transcription in the G1 phase. The retinoblastoma protein is phosphorylated at the end of the G1 phase. DNA synthesis starts after the repressive effect of RRB1 is removed and the E2F-dependent genes are activated in the S phase of the cell cycle. The Nod factor activates mitotically a cluster of cortex cells, which form the nodule primordium, and completely

differentiated cells become de-differentiated. Obviously, this is related to changes in the level of endogenous phytohormones and modulation of tissue sensibility to these phytohormones.

Arrest of the cell cycle at the G2 stage and cell endoreduplication. In the G2/M transition, a cyclin-dependent kinase/cyclin complex is formed, which is also known as a mitotic promoting factor (MPF). This complex, regulating the entry of the cell into mitosis, is completed with the formation of diploid cells. The MPF inhibition in a certain mitosis phase by degradation of cyclins B and reduction in kinase activity leads to the cell departure from mitosis at the anaphase stage and its return to the endoreduplication cycle. The anaphase-promoting complex (APC), which is activated during the G2/M phase, can be such a specific MPF inhibitor. The CCS52 protein is involved in activation of the APC complex. However, the components of signal-transduction pathways leading to cell de-differentiation and the formation of main tissues of the nodule still remain unknown.

Discussion

Analysis of fragments of the gene network involved in the early stages of interaction between the plant and bacteria strongly suggests that functioning of genes of both symbiotic partners is coordinately controlled, which leads to successful nodulation. Symbiosis is formed when genes of both partners are integrated into a superorganic system—a hybrid gene network. The formation of this type of networks is accompanied by merging of signal pathways. At early nodulation stages, signals are exchanged and a common transduction pathway is formed.

The following stage involves the delivery of bacteria into the cytoplasm of nodule primordial cells via endocytosis and their transformation to bacteroids. After this, the nodule primordium is differentiated to form a mature nodule, whose cells experience expression of plant late nodulin genes. During symbiosis, metabolic pathways of two organisms are integrated. Necessary components for fixing nitrogen in air are the expression products of these nodulin genes, products of the bacterial gene *nif* coding nitrogenase, and *fix* genes. An excess of ammonium, the final product of nitrogen fixation, inhibits the activity of the bacterial *nod* genes, which results in disruption of the gene network.

Thus, this network is an example of functioning hybrid gene networks. Data accumulation will promote reconstruction of the gene network of nodule formation and allow a mathematical model for symbiosis to be developed. Currently, the authors work on the extension of this gene network, collection of quantitative data, and construction of a gene network for a nitrogen-fixing nodule.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002) GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398-401.
2. Schultze M., Kondorosi A. (1998) Regulation of symbiotic root nodule development. Ann. Rev. Genet. 32:33-57.

GENE NETWORKS: PRINCIPLES OF ORGANIZATION AND MECHANISMS OF OPERATION AND INTEGRATION

Stepanenko I.L., Podkolodnaya O.A., Kolchanov N.A. *

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

Keywords: *gene expression, gene networks, regulatory mechanisms, integration*

Resume

Motivations: One of the major goals of the postgenomic systemic computational biology is the research into principle of organization of the gene networks controlling molecular genetic, biochemical, physiological, morphological, and other characteristics of organisms and mechanisms of their operation using the information encoded in the genomes of these organisms.
Results: Principles of organization of the gene networks described in the GeneNet database and mechanisms underlying their operation and integration are considered.

Introduction

All the processes occurring in organisms require a concerted expression of certain gene groups, forming gene networks. Modern experimental approaches result in an impetuous accumulation of huge amounts of information on the structure–function genomics, transcriptomics, proteomics, and metabolomics, allowing various organizational levels of gene networks to be comprehensively described in computerized databases. The GeneNet database, which we are developing, allows diverse data on gene network operation to be accumulated and visualized as interactive graphical layouts and contains by now description of the 25 gene network of humans, animals, and plants (Kolchanov et al., 2000; Ananko et al., 2002), controlling processes of basic metabolism, morphogenesis, response to adverse environmental factors, and others. This information is used here to consider the principles underlying gene network organization and mechanisms of their operation.

Materials and Methods

The descriptions of gene networks compiled in the GeneNet database (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet) are used in the work. Gene networks were visualized using the specialized software GeneNet Viewer. The SRS version of GeneNet (Ananko et al., 2002) allows the complete lists of elementary structures, reactions, and regulatory events specific of each gene network to be formed and the information on the interaction of a particular protein within a gene network, the genes whose transcription is regulated by a transcription factor of interest, etc., to be obtained.

Results and Discussion

Gene networks: elementary structures and events. Let us consider the structure–function organization of gene networks in the context of chemical kinetic approach divide their major components into two basic classes: (i) elementary structures (genes, RNAs, proteins, signal molecules, and various types of metabolites) and (ii) elementary events of two types—metabolic and regulatory. The elementary events comprise basic genetic processes (replication, transcription, splicing, and translation), posttranslational degradation of proteins, formation and degradation of protein complexes, biochemical reactions, transport processes, etc., and regulation of all these processes. Thus, the gene networks are described and studied at the level of elementary molecular and molecular genetic structures and their interactions using gene network graphs. A limited set of elementary structures, events, and processes forms the basis of gene network organization. However, their combinations generate a great diversity of gene networks and modes of their operation.
Compartmentalization of gene networks. Depending on specific features of organization, a certain set of compartments is involved in the function of particular gene network. The corresponding elementary structures and processes are distributed between such compartments as the nucleus, cells, cytoplasm, organelles, cell membrane, tissues, organs, and the overall body.
Gene networks: classification. Of all the possible classifications of gene networks with reference to their organization and function, let us consider the following.
1) The classification according to dynamics of the processes they control (Fig. 1): (a) gene network controlling cell differentiation and morphogenesis of tissues and organs (a monotonic drift of a parameter from the current state); (b) gene networks of homeostasis (constancy of a parameter under control); (c) gene networks of stress response (a pronounced

deviation of a parameter followed by its restoration to the initial state); (d) gene networks of cyclic processes (oscillation of a parameter); and

2) The classification according to pattern of intergenomic interactions: (i) monogenomic, involving interactions of genes belonging to the same genome, and (ii) hybrid, involving interactions of genes belonging to two or more genomes (nuclear + mitochondrial, nuclear + bacterial, nuclear + viral, etc.). The gene networks of nitrogen-fixing nodules (Ibragimova et al., 2002), the gene networks appearing upon infection of a body with bacteria or viruses, etc., belong to the latter type.

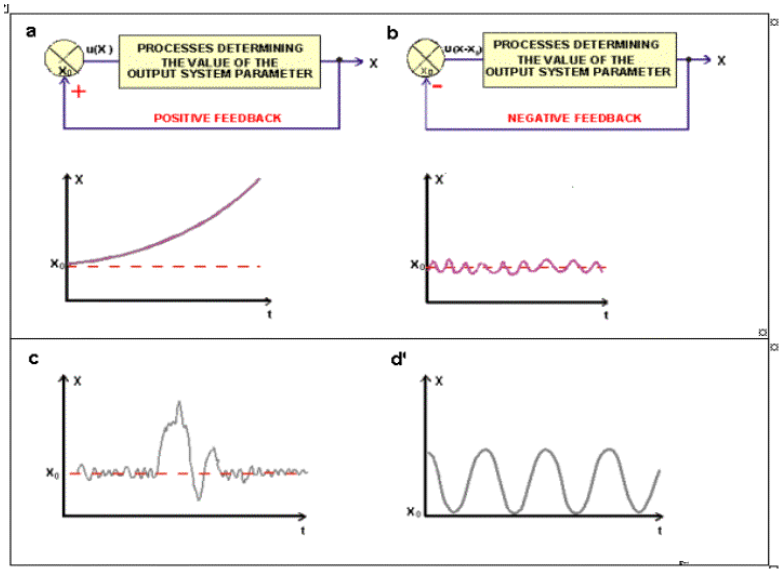


Fig. 1. Types of dynamics of the processes regulated by gene networks.

Negative feedbacks. Regulatory circuits with a positive feedback (Fig. 1b), shifting the parameters from the current state and thereby switching the gene network to a new state, as for a negative feedback (Fig. 1a), it stabilizes the parameters of gene network at a certain level. For example, the positive feedback circuits control flower development (Omelyanchuk, Aksenovich, 2002) and programmed cell death (Stepanenko, Grigor'ev, 2002), whereas negative feedback circuits control glutathione homeostasis (Kudryavtseva, Stepanenko, 2002) and synthesis of thyroid hormones (Suslov, Ignat'eva, 2002).

Signal transduction pathways. Signal transduction pathways provide for communications between the elements and compartments comprising gene networks. Each gene network has a pathway for transducing the signals from outside the cell into it. Signal transduction pathways may be characterized according to *the type of external signal*—a light signal (Smirnova et al., 2002), protein, nonprotein signal molecule (steroid hormones, amino acid derivatives, NO), etc.; *the type of receptor* that receives the external signal—transmembrane or intracellular receptor; to the *internal links* dependent on the type of receptor (a cascade of protein kinase reactions, a network of proteolytic reactions, etc.); *terminal link* that completes the signal transduction pathway—a nuclear gene, gene in an organelle genome, RNA, or cytoplasmic protein; or to *molecular product of the terminal link*—DNA–protein complex between a transcription factor and its binding site, DNA–protein complex formed by two transcription factors within a composite element or active oxygen species. Neurohumoral signal transduction pathways play an important role in the supracellular gene networks.

Central regulators of gene networks. As a rule, each gene network has one central regulator—a protein providing the coordination of numerous gene network elements. The central gene network regulator in the majority of the cases is a transcription factor. An increase in its concentration in the cell nucleus triggers a ***cassette transcription activation*** of many genes through interaction of this factor with its binding sites in promoters of the corresponding genes. Shown in Fig. 2 are the central regulators and the gene cassettes they activate for several gene networks, including those controlling the cell cycle, cholesterol biosynthesis, erythrocyte differentiation, heat shock response, etc.

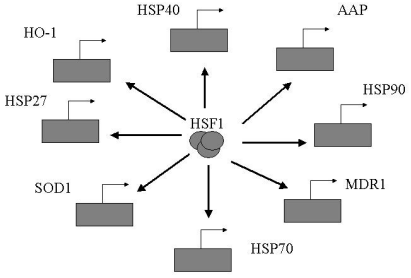


Fig. 2. A cassette-type transcription activation by the central network regulator by the example of the gene network controlling heat shock response.

The mechanisms activating central regulators involve increases in the level of the signals received. Positive feedbacks play the role of the utmost importance in increasing the signals (Fig. 3). These increases may involve (A) protein–protein interactions (a caspase cascade); (B) self-enhancement of transcription of a gene encoding a transcription factor; (C) self-enhancement of a gene encoding a receptor; (D) mutual increase in the transcriptions of two genes; and (E) mutual increase in the transcriptions of two genes via a heterodimeric transcription factor. Mechanisms allowing the signal to be amplified at the gene regulatory level may involve (F) synergistic action of transcription factors at the level of composite elements.

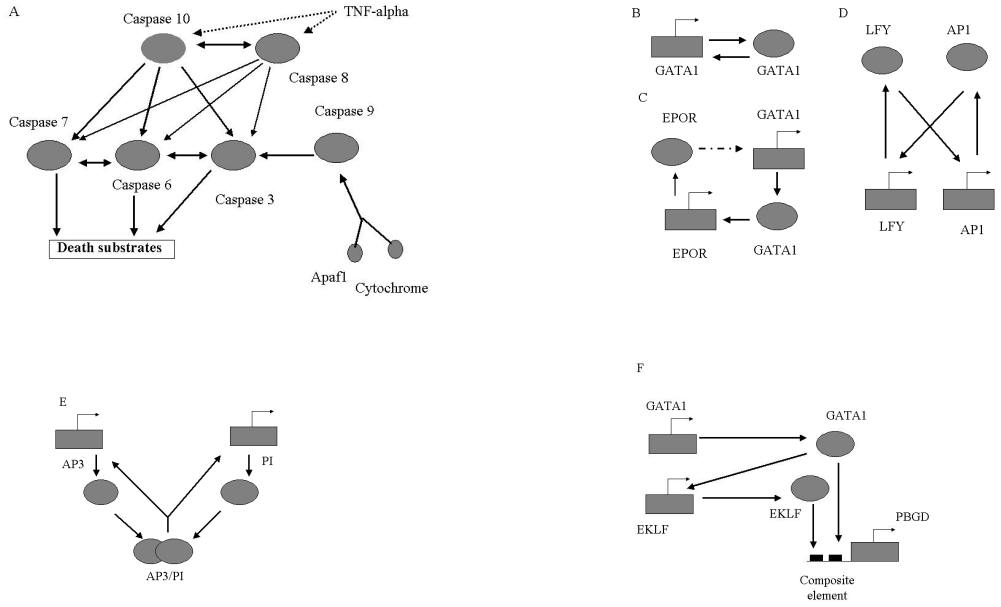


Fig. 3. Increase in the signal of the central regulator by positive feedback mechanisms (see text for details).

Interaction of regulatory circuits of gene networks. Typical of gene networks is interaction of regulatory circuits providing the diversity of their operation modes. Let us consider a number of examples.

I) Gene networks of homeostasis (linking of negative feedbacks). In the gene network controlling cholesterol biosynthesis (with the transcription factor SREBP as a central regulator), the intensity of cholesterol biosynthesis in the mevalonate pathway is under the control of *a negative feedback,* while the intensity of low density lipoprotein transport into the cell, by another negative feedback, thereby providing a finer tuning on the blood cholesterol level (Fig. 4).

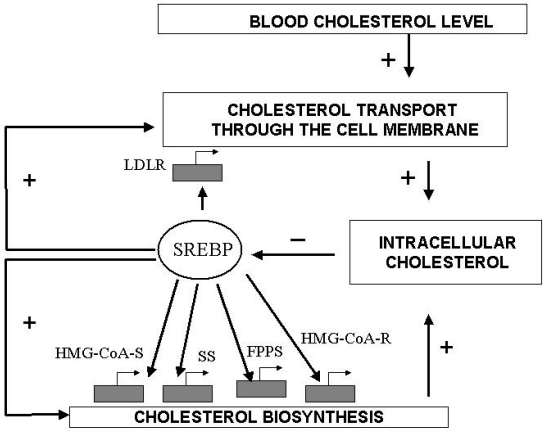


Fig. 4. Linking of two negative feedback circuits (gene network of cholesterol biosynthesis).

II) Cyclic gene networks (interaction of positive and negative feedbacks). In the gene network controlling the cell cycle (Fig. 5a), the central regulator during G0 phase—the transcription factor E2F1 within the trimer E2F1/DP1/pRB—represses a cassette of genes (Turnaev, Podkolodnaya, 2002). The positive feedback circuits underlie derepression of these

genes in G1 phase and their activation by the dimer E2F1/DP1. The negative feedbacks switched on thereafter inhibit the transcription of this gene cassette with the trimer E2F1/DP1/pRB and finally return the gene network of cell cycle regulation to its initial state (early G1 phase). Thus, the interaction of these regulatory circuits triggers the gene network of cell cycle between different states.

III) Gene networks of stress response (interaction of positive and negative feedbacks). In the gene network of antiviral response (Fig. 5b), a positive feedback initially activates transcription of the gene encoding the transcription factor IRF-1, which, in turn, activates the transcription of interferon beta gene. Later, the factor IRF-1 activates the transcription of another factor—IRF-2—thereby restoring the initial state of the gene network. Return of the stress response gene networks into the quiescent state may be provided by another mechanisms, for example, inactivation of a transcription factor through its binding to the inhibitor whose gene is activated by this transcription factor (the gene network of heat shock response; Fig. 5c).

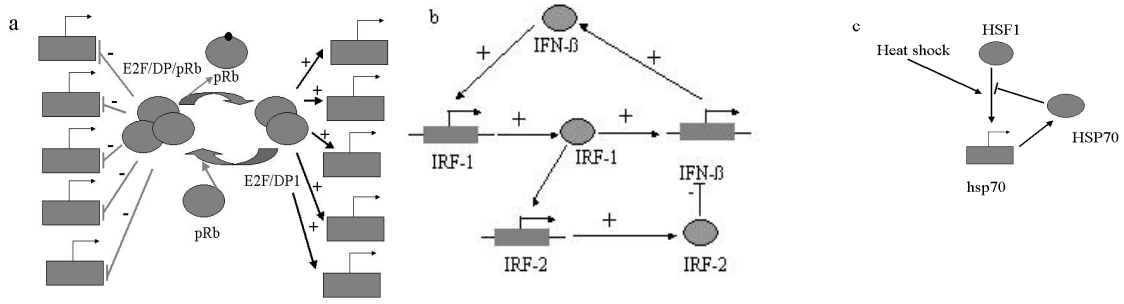


Fig. 5. Interaction of positive and negative feedbacks: (**a**) gene network of cell cycle regulation and (**b, c**) gene network of stress response (see text for details).

IV) Gene networks controlling morphogenesis (interaction of negative regulators with positive feedback circuits). In the gene network of floral development, the signal transduction pathway via the transcription factor CO activates transcription of the gene TFL1 under conditions of a short day. The corresponding protein TFL1 represses the genes AP1 and LFY, halting the meristem development at the vegetative stage (Fig. 6). However, under conditions of a long day, transduction of signal via the same factor CO activates the gene LFY, whose product activates transcription of the gene AP1. In turn, the protein AP1 activated the gene LFY, thereby closing the positive feedback circuit with two mutually activated genes. When AP1 and LFY critical concentrations are reached, the gene TFL1 is inhibited, triggering the floral development processes. Thus, the trigger mechanism underlies the transition of meristem in another stable state—development of flower.

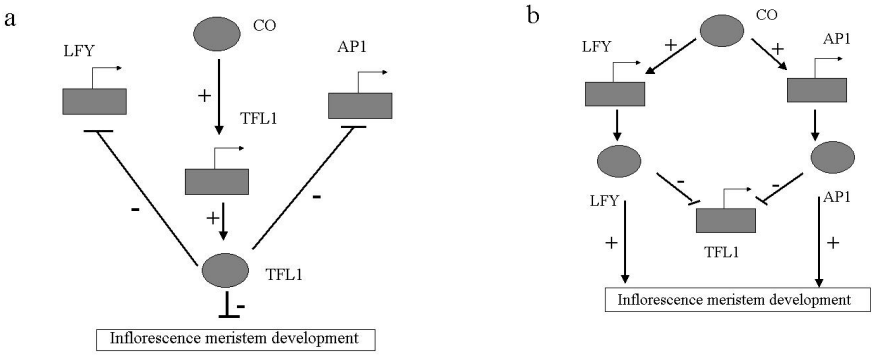


Fig. 6. Two states of gene network regulating floral developments: (**a**) short day and (**b**) long day.

Integration and interaction of gene networks. Local gene networks are integrated into the global gene network of the overall organism. The integration of gene networks requires coordination in the flows of matter, energy, and information between outputs of some gene networks and inputs of other gene networks. Neuroendocrine system plays the key role in integration of the local gene networks at the level of the overall organism. For example, three hierarchical levels are apparent in the gene network of thyroid system (Suslov, Ignatieva, 2002); they correspond to (i) thyroid gland, producing thyroid hormone; (ii) hypophysis, secreting thyrotropin (TSH); and (iii) hypothalamus, producing thyrotropin-releasing hormone (TRH). Separate gene networks, integrated through neurohormonal signals, operate at each of these three levels. Totally, this gene network comprises two positive and six negative feedback circuits. ***Integrator gene networks*** play the

key role in integration of local gene networks at the level of individual cell. Let us consider the gene network of redox regulation , providing the adaptation to an oxidative stress, as an example (Fig. 7).

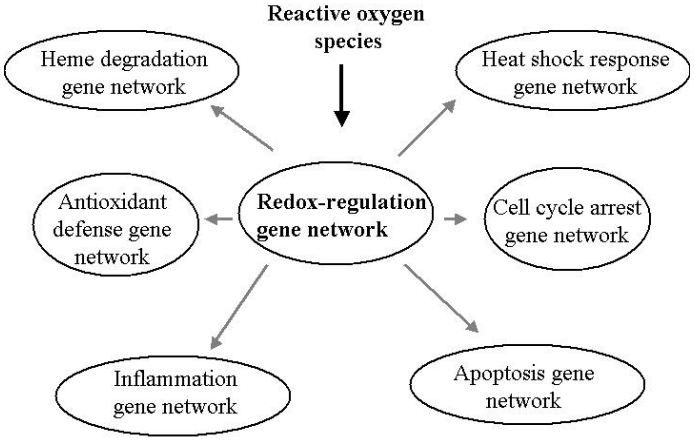


Fig. 7. Integration of gene networks.

The following cassette of gene networks is switched on when the gene network of redox regulation is switched on: (1) gene network of antioxidant defense, (2) of cell cycle arrest, (3) inflammation, (4) iron catabolism, (5) heat shock response, (6) gene network of apoptosis. Thus, the integrator gene network, receiving a certain signal through the input reception system, processes this signal and distributes it between its outputs to activate the connected gene networks; each of these gene networks may, in turn, transduce the activating signal to one or several other gene networks. Actually, the integrator gene networks are the key elements providing transduction of an activating signal through the nodes of the global gene network of the overall organism.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376, 01-07-90084, 00-07-90337, 02-07-90355, 02-07-90359, 00-04-49229, and 00-04-49255); Ministry of Industry, Science, and Technology of the Russian Federation (grant № 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Projects № 65 and 66); US National Institutes of Health (grant № 2 R01-HG-01539-04A2); and US Department of Energy (grant № 535228 CFDA 81.049). The authors thank I.V.Lokhova and L.V.Katokhina for their assistance with bibliography.

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002). GeneNet: a database on structure and functional organisation of gene networks. Nucl. Acids Res. 30:398-401.
2. Ibragimova S.S., Smirnova O.G., Shavrukov Yu.N., Stepanenko I.L. (2002). A hybrid network of nitrogen-fixing nodules: intergenomic interactions of bacteria and host plant. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
3. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000). Gene networks. Mol. Biol. (Mosk). 34:533-544.
4. Kudryavtseva A.N., Stepanenko I.L. (2002). Gene network of glutathione homeostasis: a response to oxidative stress. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
5. Omelyanchuk N.A., Aksenovich A.V. (2002). Fragments of gene network of flower development in Arabidopsis under long day conditions and their description in the GeneNet system. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
6. Smirnova O.G., Ibragimova S.S., Shavrukov Yu.N., Stepanenko I.L. (2002). Negative regulation of plant photomorphogenesis. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
7. Stepanenko I.L., Grigor'ev S.A. (2002). Organization of the gene network of apoptosis. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
8. Suslov V.V., Ignat'eva E.V. (2002). Molecular genetic mechanisms regulating the thyroid system: description in the TRRD and GeneNet databases. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
9. Turnaev I.I., Podkolodnaya O.A. (2002). Gene network on cell cycle control. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

MEASUREMENTS OF PRECISION OF MOLECULAR MECHANISMS FOR EARLY *DROSOPHILA* EMBRYO SEGMENTATION

Spirov A.V. [1,2]***, Holloway D.M.*** [3]

[1] State University of New York at Stony Brook, Stony Brook NY 11794-3600
[2] The Sechenov Institute of Evolutionary Physiology and Biochemistry, 194223, St.-Petersburg, Russia
e-mail: spirov@kruppel.ams.sunysb.edu
[3] Mathematics/B.C.Institute of Technology/Burnaby, B.C. Canada, e-mail: David_Holloway@bcit.ca

Key words *: early development, segmentation mechanisms, expression profiles, embryo-to-embryo variability, nucleus-to-nucleus variability, morphogenetic gradients, gradient reading, positional information, patterning precision*

Resume

Motivation: Quantitative estimation of precision of early Drosophila embryo expression patterning in terms of embryo-to-embryo and nucleus-to-nucleus variability give the possibility of treating classic problems of developmental regulation in terms of modern functional genomics.
Results: We confirm the recent findings of Houchmandzadeh with co-authors (2002) concerning high embryo-to-embryo variability of the maternal morphogen *bicoid (bcd)* with comparably lower variability of the gap gene product *hunchback (hb)*. We also present data showing relatively high variability in the posterior maternal morphogen *caudal (cad)* with lower variability in pair-rule product *even-skipped (eve)*.

Introduction

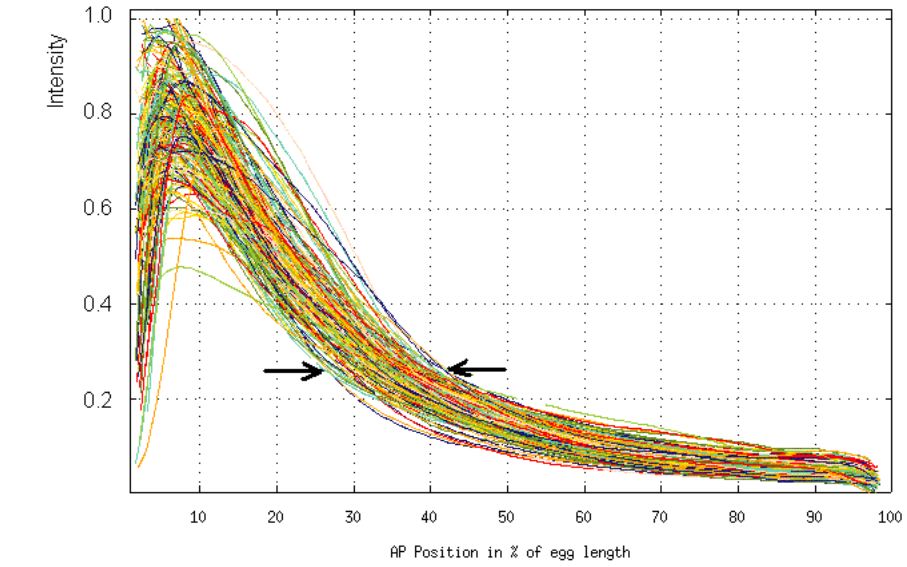
Though much progress has been made on the molecular characterization of developmental gradients in the century since they were initially proposed, fundamental questions remain as to robustness of gradients for specifying positional information in embryonic development.
These are:

1. Size regulation, or scaling, i.e., pattern remains unaltered, despite high variability in embryo size.
2. Gradient stability in the face of fluctuations in temperature and other environmental factors.
3. Transmission of positional information by gradient concentration, in the presence of inherent (due to diffusion and reaction) concentration fluctuations.

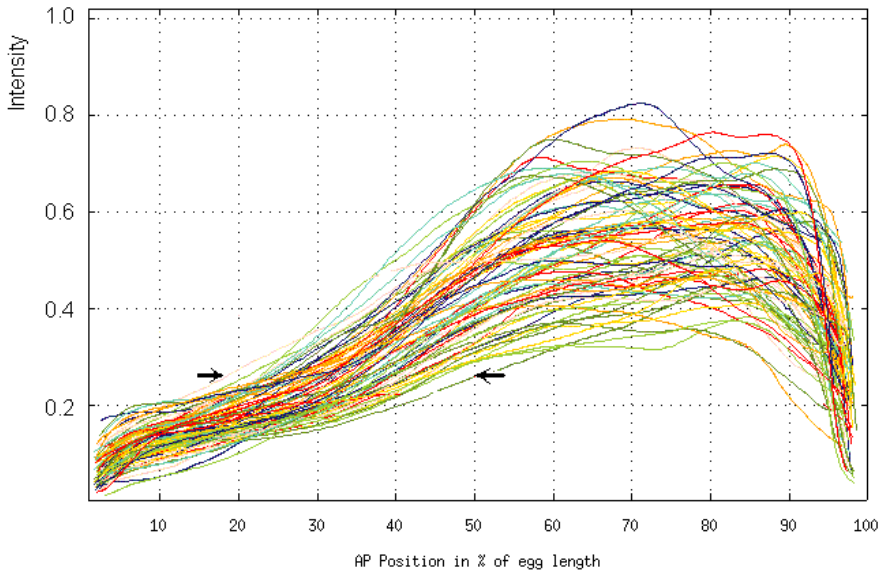
This problem includes the transmission of errors down signaling cascades. Segment determination in the fruit fly *Drosophila melanogaster* is one of the best characterized (molecularly) systems for studying developmental pattern formation. New techniques have allowed us to take some steps towards quantifying variability in this system, shedding light on some very old embryological questions.

Methods and Algorithms

Images of *Drosophila* Gene Expression. Processing of embryonic images begins with data expressed in terms of the average fluorescence level (proportional to gene expression level) at 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 et al., 1998). These images were computationally treated by means of the *Khoros* package (Rasure, Young, 1992). Embryos were rotated and cropped automatically. Next, the images were *segmented* (Kosman et al., 1997). 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.
Anterior-Posterior Expression Profiles. Because the expression of segmentation genes is largely a function of position along the anterior-posterior (A-P) axis, it is natural to use the A-P profiles of gene expression as a first step towards characterization of embryo-to-embryo variability. We use expression data from a strip (of width 10% Dorsal-Ventral, D-V height) along the midline of an embryo in the A-P direction. The D-V values of the data are then ignored and fluorescence intensity (y) is plotted against A-P direction (x).



A



B

Fig. 1. High variability of anterior and posterior primary morphogens in the early *Drosophila* embryo. A. *Bicoid* gradient of about 300 cycle 14 embryos. B. *Caudal* gradient of about 100 cycle 14 embryos.

Temporal Classification. All the embryos under study belong to cleavage cycle 14. We divided the embryos into temporal classes by an extensive and thorough visual analysis of images and graphs of individual embryos (Myasnikova et al., 2001). Each image is allocated to one of 8 temporal classes on the basis of visual inspection of the (highly dynamic) expression pattern of the pair-rule gene *eve*.

Results and Discussion

Following Houchmandzadeh et al. (2002) we use a simple but illustrative way of presenting embryo-to-embryo variation in expression profiles. Figure 1 presents about 300 profiles of the anterior maternal factor *bcd* and about 100 profiles of posterior maternal factor *cad*. All profiles were normalized for embryo length. We observe that both *bcd* and *cad* display high levels of embryo-to-embryo variability. The positions at which the *bcd* profiles cross a chosen intensity level (23%) spread over about 20% of the embryo length (Fig. 1A). Houchmandzadeh et al. (2002), observed approximately 30% variability in crossing this intensity threshold, using about 100 *bcd* profiles. The variability in *cad* profiles is also very high (Fig. 1B).

Embryo-to-embryo variability of the primary pair-rule factor *EVEN-SKIPPED (eve)* appears to be much less than for the maternal factors. Variability in crossing any given threshold on the anterior or posterior slopes of any of the seven expression peaks appears to be in the range of 5-7% embryo length (Fig. 2). This result is comparable to the 4% variability found by Houchmandzadeh et al. (2002) at a mid-embryo position for the gap gene *hunchback (hb)*, indicating that maternal factor variability is decreased as it is transmitted down the segmentation cascade.

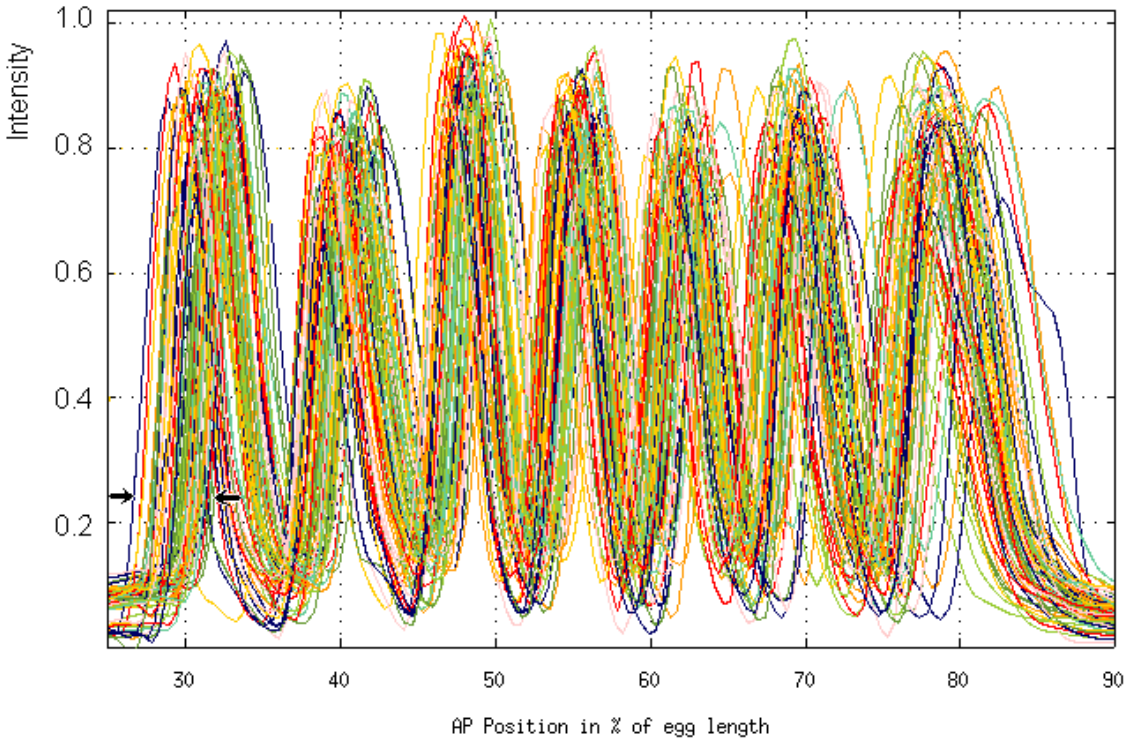


Fig. 2. Variability of profile of expression for primary pair-rule gene *even-skipped* in the early Drosophila embryo (about 100 profiles of late cycle 14 embryos).

If no error suppression is occurring in reading upstream positional information, theory and computations (Lacalli, Harrison, 1991; Holloway, Harrison, 1999) predict an increase in positional errors going down the segmentation hierarchy. Our experimental results indicate that error suppression is a component of the molecular mechanism for segmentation patterning, favoring proposed mechanisms with the ability to filter noise over those which rely on simple gradient cueing.

Acknowledgements

This work (AVS) is supported by USA National Institutes of Health grant RO1-RR07801; INTAS grant № 97-30950 and RFBR grant № 00-04-48515.

References

1. Houchmandzadeh B., Weischaus E., Leibler E. (2002) Establishment of developmental precision and proportions in the early Drosophila embryo. Nature. 415: 798-802.
2. Holloway D.M., Harrison L.G. (1999). Suppression of positional errors in biological development. Mathematical Biosciences. 156: 271-290.
3. Kosman D., Small S., Reinitz J. (1998). Rapid preparation of a panel of polyclonal antibodies to Drosophila segmentation proteins. Development, Genes, and Evolution. 208: 290-294.
4. Kosman D., Reinitz J., Sharp D.H. (1997). Automated assay of gene expression at cellular resolution. In Altman R., Dunker K., Hunter L., Klein T. editors. Proc. of the 1998 Pacific Symposium on Biocomputing. Singapore: World Scientific Press. 6-17.
5. Lacalli T.C., Harrison L.G. (1991). From gradients to segments: models for pattern formationin early *Drosophila* embryogenesis. Seminars in Developmental Biology. 2: 107-117.
6. Myasnikova E.M., Samsonova A.A, Kozlov K.N, Samsonova M.G., Reinitz J. (2001) Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. Bioinformatics. 17: 3-12.
7. Rasure J.R., Young M. (1992) An Open Environment for Image Processing Software Development. Proc. SPIE. 1659: 300-310.

BIOINFORMATIC ANALYSIS OF A MORPHOGENETIC FIELD IN *DROSOPHILA*

Reinitz J.B.

Dept. Of Applied Math and Statistics, Stony Brook University, Stony Brook, New York 11794-3600 USA
e-mail: Reinitz@ams.sunysb.edu

Key words: *Drosophila, development, transcription, databases, optimization, dynamical model, gene expression, reverse engineering, knowledge representation, theory*

Resume

Motivation: To construct a physico-chemical model of a morphogenetic field in Drosophila.
Results: The project is ongoing. We have achieved biologically significant results with the model. Other results include a new database of gene expression and new optimization methods.
Availability: Many of the pieces of software used in these studies are available by request from the author.

Introduction

The impact of genomics on biology is having revolutionary results. As the use of "functional genomics" expands from low level function like enzymatic specificity to higher level functions in immunology and neurobiology, the ultimate outcome will be the integration of genomics into the mainstream of biology. The fundamental strength of the genomics approach is its ability to identify a complete set of active biological players, where the categories involved—genes, transcripts, protein structures–depend on the specific application. For fundamental studies in development, the critical entity is a morphogenetic field, and the complete set of players is the full set of chemical regulators of that field.
We are engaged in a comprehensive study of the morphogenetic field controlling Drosophila segmentation. Our approach is based on an integrated program of theory and experiment. The theoretical component uses ordinary and partial differential equations together with large scale computation to achieve more powerful ways to organize and interpret the data, while the experimental part employs automated image processing and data analysis methods for the construction of an integrated spatio-temporal map of gene expression at cellular resolution. This program takes advantage of special properties of the Drosophila segmentation system, as we now explain.

Methods and Algorithms

I. The Biological System
Immediately following fertilization and egg deposition, the newly formed zygotic nucleus undergoes a series of rapid and synchronous nuclear divisions. By the ninth such division the nuclei have migrated to the cortex (outside) of the egg, and the embryo begins a stage of development called the syncytial blastoderm. Although cell membranes do not form, each nucleus is surrounded by an island of cytoplasm, called an energid, stabilized by cytoskeleton which is separated from other energids by noncytoplasmic material. Hence the syncytial blastoderm differs from a classical syncytium like a muscle fiber, and may be thought of as a collection of cells not delimited by membranes. In the last half hour of the blastoderm stage cell membranes invaginate between the blastoderm nuclei, sealing them off into cells. When cellularization is complete, gastrulation begins.
The classical genetics of segmentation is well characterized. A significant feature of Drosophila is that it is often possible to obtain a complete set of genes affecting a particular function by saturation mutagenesis. This key feature of the system was used to identify and clone the segmentation genes. These genes are grouped into four classes: coordinate genes (expressed from the maternal genome), gap genes, pair-rule genes, and segment polarity genes. Gap and pair-rule genes are expressed beginning at the onset of syncytial blastoderm. Gap genes assume characteristic patterned expression immediately, while pair-rule genes do not do so until cleavage cycle 14. Segment polarity genes are not expressed until about the time of gastrulation. With respect to spatial expression, gap genes are typically expressed in two domains 10-20 nuclei wide, pair-rule genes in 7 stripes, each 3-6 nuclei wide, and segment polarity genes in 14-17 stripes one nucleus wide.
II. Scientific Approach
In a typical developmental process, well characterized genetics alone does not provide enough information to physiologically model the process. The blastoderm is a very important exception to this generalization. First, the blastoderm is a syncytium, so that cell-cell signaling can be neglected. Spatial interactions can be treated in terms of diffusion of gene products between blastoderm nuclei. Second, segment determination happens because of differential gene expression, which is directly observable. The segmentation genes have been cloned, and hence their level of expression can be monitored by antibody or hybridization methods. Third, the system of segmentation genes is not coupled to other developmental processes until after gastrulation, because morphological alterations due to mutations in

segmentation genes are not apparent until after gastrulation. The second and third reasons together mean that all the state variables of this system are directly observable. This is a very unusual situation in biology: It affords an opportunity to understand important aspects of developmental genetics in unprecedented detail.

We are making an effort to analyze this system at two simultaneous biological levels, pattern formation and the control of transcription. Pattern formation in the segmentation morphogenetic field involves the interaction of many identical genetic networks, each in a cell nucleus. In our study of pattern formation, we do not model the molecular substructure of transcriptional control explicitly because understanding the molecular hardware of transcriptional control is a difficult problem in its own right. We are approaching the transcription problem by modelling embryos transformed with certain lacZ constructs containing well characterized fragments of of segmentation gene promoters. Our long term goal is to obtain the pattern formation equations as a coarse-grained approximation to the transcription equations.

Our approach to both of these problems has a common framework and consists of 4 components: (1) The formulation of a **theoretical model** for gene regulation at a given level of description. (2) The acquisition of **gene expression data** using fluorescently tagged antibodies and/or RNA probes. (3) The determination of the values of parameters in the model or the demonstration that no such values exist by **numerical fits to data**. 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. We very briefly outline these areas below.

Implementation and Results

The Theoretical Model for Pattern Formation

The expression of segmentation genes is a function of their position along the anterior-posterior (A-P) axis, and so we work with a 1D model. Let the position of a cell nucleus along the A-P axis be indexed by i, such that nucleus $i+1$ is immediately posterior to nucleus i. Each cell nucleus contains a copy of a regulatory circuit composed of N genes, determined by an $N \times N$ matrix **T**. The concentration of the a th gene product in nucleus i is a function of time, denoted by $v_i^a(t)$. Then

$$\frac{dv_i^a}{dt} = R_a g_a\left(\sum_{b=1}^{N} T^{ab} v_i^b + m^a v_i^{bcd} + h^a\right) + D^a(n)\left[\left(v_{i-1}^a - v_i^a\right) + \left(v_{i+1}^a - v_i^a\right)\right] - \lambda_a v_i^a, \qquad (1)$$

where N is the number of genes included in the circuit. The first term on the right hand side of the equation describes gene regulation and protein synthesis, the second describes exchange of gene products between neighboring cell nuclei, and the third represents the decay of gene products. g_a is a "regulation-expression function", which we assume takes the form $g_a(u_i^a) = (1/2)[(u_i^a/\sqrt{u_i^{a\,2}+1})+1]$ for all a, where $u_i^a = \sum_{b=1}^{N} T^{ab} v_i^b + m^a v_i^{bcd} + h^a$.

In (1), T^{ab} is the regulatory matrix. The elements of this matrix may be thought of as sensitivity coefficients, since a small change in the concentration of a regulator v^b will produce a change in the synthesis rate of its target v^a that is at most proportional to T^{ab}. The bcd input is given by $m^a v_i^{bcd}$, where v_i^{bcd} is the concentration of bcd protein in nucleus i and m^a is the regulatory coefficient of bcd acting on zygotic gene a. R_a is the maximum rate of synthesis from gene a, and h^a summarizes the effect of general transcription factors on gene a. The diffusion parameter $D^a(n)$ depends on the number n of cell divisions that have taken place, and varies inversely with the square of the distance between nuclei. We assume that the distance between adjacent nuclei is halved after a nuclear division. λ_a is the decay rate of the product of gene a. This equation can also cast into the form of a set of partial differential equations (PDE's) [3], an approach that confers certain advantages in the mathematical analysis.

The Theoretical Model for Transcription

We have also developed new equations for transcription designed to elucidate the rules by which interactions among bound ligands give rise to modular enhancers and sets of modular enhancers give rise to the behavior of intact genes. These equations take the form of complex feedforward functions which have explicit representations of a variety of proposed mechanisms of activation and repression which act at different scales. Although not all mechanisms to be represented have as yet been incorporated in these equations, they are now sufficiently well developed to begin initial comparisons with data.

Experimental Data

All of the parameters in the pattern formation and transcription equations will be determined by fits to data. Because the trans-regulators of the constructs we model are the segmentation gene products themselves, the data used for both models is very closely related. We construct our numerical dataset by a seven step procedure: 1) Embryos are stained with antibodies (for proteins) and/or hybridized to lacZ probe. 2) The embryos are confocally scanned. 3) Images are segmented. 4) Each embryo is placed in a temporal equivalence class. 5) The embryos are registered on a common expression domain (Myasnikova et al., 2001). 6) Background staining is numerically removed. 7) Data from the same

genes is averaged. This data has been placed in the database FlyEx, accessable on the worldwide web at http://www.csa.ru/flyex. or at http://flyex.ams.sunysb/FlyEx.

Fits to Data

Circuit parameters are determined by least squares fits to gene expression data using the method of simulated annealing. I will discuss our current algorithms for serial and parallel simulated annealing. We have developed a new method for statistically tuned parallel simulated annealing, based on a genetic algorithm-like selection scheme. The algorithm performs well as long as the selection is performed while the processors are statistically independent but close to thermal equilibrium, and this tuning can now be achieved by means of new statistical estimators.

Biological Validation

The pattern formation model has achieved some noteworthy results with much cruder data than is reported here. We are finally able to report the initial results of fits to the new dataset, which include a number of interesting biological features. Any initial results of fits to the transcriptional model will be reported.

Discussion

This project for the characterization of the segmentation morphogenetic field has been ongoing for 13 years. Because of the intellectual contributions of many coworkers, who will be mentioned by name in my talk, we are now in reach of a true physiological model of a morphogenetic field. The preliminary results reported in this talk indicate that such a model, together with its bioinformatics support, is likely to be helpful in understanding important biological questions.

Acknowledgements

This work was supported by grants RR07801 and TW01147 from the US NIH. The author thanks his coworkers D.Sharp, M.Samsonova, A.Samsonov, E.Myasnikova, S.Hou, S.Surkova, K.Kozlov, V.Gursky, A.Pisarev, K.P'ustelnikova, C.Alonso-Vanario, D.Kosman, J.Jaeger, S.Spirov, Y.Wang, K.Chu, L.Greenwald, L.Carey, and Manu.

References

1. Gursky V., Reinitz J., Samsonov A. (2001). How gap genes make their domains: an analytical study based on data driven approximations. Chaos. 17:3–12.
2. Myasnikova E., Samsonova A., Kozlov K., Samsonova M., Reinitz J. (2001). Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods. Bioinformatics. 17:3–12.

TEMPORAL CHANGES IN POSITION OF SEGMENTATION GENE EXPRESSION DOMAINS IN *DROSOPHILA* EARLY EMBRYO

**** Surkova S.Yu., Samsonova M.G., Myasnikova E.M.***

St.-Petersburg State Technical University, Russia, e-mail: surkova@fn.csa.ru
*Corresponding author

Key words: *Drosophila, segmentation, gene expression, positional information, statistical analysis*

Resume

Motivation: Expression of segmentation genes plays a crucial role in organization of the *Drosophila* segmented body plan. We investigate the mechanisms of segment determination by the integrated program of mathematical modelling and experiment. Recently we have acquired large amount of quantitative data on expression of segmentation genes at cellular resolution. The unprecedented quality of this data makes possible to find new biological relations, which could be used for validation of the model. One way to reveal these relations is statistical analysis of the data that will be considered in this paper.
Results: The *x* coordinates of extrema values were considered as characteristic features of expression domains for 9 segmentation genes. Analysis of changes in position of these domains was performed both for 8 discrete time points (time classes) and for a continuous developmental period in minutes. We have shown that most of expression domains significantly change their position with time. The dynamical shifts in location of pair-rule gene expression domains depend considerably on formation of new domains during development.
Availability: All data are available from authors.

Introduction

One of the basic problems in developmental biology is a mechanism of pattern formation. Pattern formation is a way in which a single cell (zygote) gives rise to the gradual formation of a set of different cell types arranged in a specific order, which is characteristic for each species. One of the primary patterning decisions required in fruit fly *Drosophila* is the division of a major body axis into serially repeated units or segments. The determination of segments (i.e., stable specification of the developmental fate) in *Drosophila* is thought to be a consequence of the expression of segmentation genes. The classical genetics of segmentation is well characterized (Nüsslein-Volhard et al., 1985). Of particular importance are members of the "gap" and "pair-rule" classes of segmentation genes. Gap genes are expressed in one to three domains, while pair-rule genes initially express protein in a single very broad domain that restricts to seven narrow domains (stripes) over a relatively short time interval. The striped pair-rule expression patterns are the first direct manifestation of the periodic pattern of the segments.
We investigate the mechanisms of segment determination by the integrated program of mathematical modelling and experiment. Mathematical modelling is based on method known as gene circuits (Reinitz et al., 1995; 1998), while experimental work is performed to acquire data on segmentation gene expression at cellular resolution (Kozlov et al., 2000; Myasnikova et al., 2001). The unprecedented quality of this data enables to reveal new biological meaningful relations for the model validation. One way to find these relations is the statistical analysis of the data that will be considered in this paper.

Materials and Methods

We obtain images of gene expression patterns as described in (Kosman, 1997). Image processing procedures resulted in the reduction of image information to a quantitative data on gene expression. At present, our dataset contains confocal scans of about 1400 embryos, of which 809 are wild type and belong to cycle 14A. The embryos were scanned for the expression of 13 segmentation genes. These embryo images were distributed by visual inspection of pair-rule gene expression patterns into 8 temporal equivalence classes (Myasnikova et al., 2001). For 103 wild type embryos, the precise developmental time was determined by measuring the degree of membrane invagination (Merill et al., 1988). Characteristic features of 1D gene expression patterns were extracted by means of the wavelet decomposition of the signal (Kozlov et al., 2000). We use standard statistical procedures (StatSoft Statistica package) to validate the significance of positional changes in gene expression patterns. The purpose of analysis of variance (ANOVA) was to test the differences in mean positions of peaks between 8 time classes for statistical significance.

Results and Discussion

Expression of *Drosophila* segmentation genes is largely a function of the position on the anterior-posterior (A-P) axis, and so can be well represented in one dimension. The representative 1D expression patterns of pair-rule and gap genes expression are shown in Figure 1.

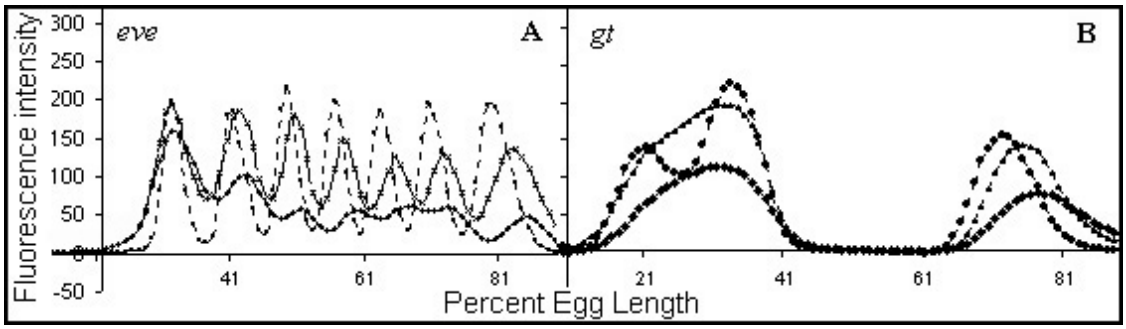


Fig. 1. Examples of expression patterns of segmentation genes. **A**. Formation of seven stripes of expression of *eve* gene, which belongs to pair-rule class. **B.** Domains of expression of *gt* gene. More sharply defined expression domains correspond to the later expression patterns.

In this study we have considered *x* positions of extrema values as characteristic features of segmentation domains. We have examined 4 gap genes: *gt, Kr, kni, hb* from temporal classes 1 to 8 and 5 pair-rule genes: *eve, ftz, h, run*, and *odd*, from temporal classes 3 to 8.
Table 1 demonstrates the significant temporal changes in position of posterior domains of gap gene expression.

Table 1. Changes in positions of gap domains*.

Gene	*Kr* {231}	*kni* {122}	*gt* {126}	*hb* {159}
tc	2-8	1-8	1-8	1-8
shift	2.66	3.20	7.57	7.49

Temporal shifts in location of the pair-rule expression domains (Table 2) appeared to be more complicated. Our observation indicates that there are two types of shifts within the pair-rule expression patterns.
The first type is caused by the formation of new peaks with time, while the second type arises due to the shift of the posterior border of expression pattern.

Table 2. Shifts in position of the expression domains of pair-rule genes.

Gene		1max	1min	2max	2min	3max	3min	4max	4min	5max	5min	6max	6min	7max
eve {654}	tc	3-8	3-8	3-8	3-8	3-8	3-8	3-8	3-8	4-8	4-8	4-8	3-8	3-8
	shift	0.47	1.87	1.94	2.77	1.97	2.27	3.44	3.02	3.21	2.54	2.62	3.82	5.258
	p<	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}
ftz {158}	tc	3-8	3-8	3-8	3-8	3-8	3-8	4-8	4-8	3-8	3-8	4-8	4-8	5-8
	shift	0.2	0.66	1.27	2.07	1.72	3.89	1.86	0.10	2.27	3.84	3.84	2.51	1.01
	***p*<**	.396	.0004	10^{-6}	10^{-6}	10^{-6}	10^{-6}	.0001	.874	10^{-6}	10^{-6}	10^{-6}	.0001	.087
h {105}	**tc**	3-8	3-8`	3-8	3-8	3-8	4-8	4-8	4-8	4-8	4-8	4-8	4-8	4-8
	shift	-2.18	0.16	1.54	2.21	3.61	2.34	1.11	2.12	2.61	2.63	3.25	3.19	3.82
	***p*<**	10^{-6}	.0496	.0005	10^{-6}	10^{-6}	10^{-6}	.0139	10^{-5}	0	10^{-6}	10^{-6}	10^{-6}	10^{-6}
odd {85}	**tc**	3-8		4-8		3-8		5-8		3-8		3-8		
	shift	2.41		1.13		2.98		0.92		3.65		2.66		
	***p*<**	10^{-6}		.6507		10^{-6}		.0321		10^{-6}		10^{-5}		
run {65}	**tc**	3-8	3-8	3-8	3-8	3-8	3-8	4-8	4-8	4-8	3-8	3-8	5-8	5-8
	shift	0.20	1.07	1.66	1.26	2.81	4.11	2.56	2.01	1.57	1.68	3.41	3.85	3.61
	***p*<**	.7382	.0125	.0004	.0421	10^{-6}	10^{-6}	10^{-6}	.0036	.0052	10^{-6}	10^{-6}	10^{-6}	10^{-6}

* In **tables 1** and **2**: **tc:** time classes for which peak positions were determined; **shift:** shift of the peak positions (% embryo length (EL)), positive- in the posterior-anterior (P-A) direction, negative- in the A-P direction; ***p***-level: significant are shift values with $p < 0.05$. Sample sizes for each gene are shown in braces. The cells are left blank, if the features could not be extracted. In **table 1** all ***p*-** values are less than 10^{-6}.

Our results demonstrate that the later formed pair-rule peaks always shift the already existing peaks or shift themselves. The mode of shift depends on how a new peak is formed. The shift of the 7th maximum (posterior border of expression pattern) is not caused by formation of new peaks in the neighborhood. This maximum moves significantly in *eve, h* and *run*, but does not move in *ftz.* The effect is a "shrinking" of patterns of the first three of above mentioned genes, while the expression pattern of *ftz* does not become more narrow with time. Unlike the posterior maximum, the first maximum moves slightly only in the cases, where some new domains are formed nearby (*h, odd*).

Most of the peaks move from posterior to anterior, or to the left in the standard orientation of the embryo. Only the peaks located in the anteriormost (head) area can move (or shift other peaks) in the opposite direction, i.e. from anterior to posterior. Besides the values of the positional shifts shown in the Tables 1 and 2, one should also take into account their standard deviations, which vary from 0.7 to 1.9% of embryo length. This positional error mainly depends on sample sizes, which are variable for different genes, as well as on the scattering of extrema positions.

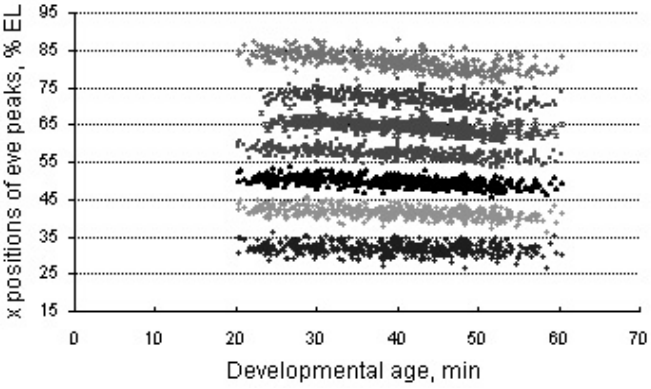


Fig. 2. Dynamics in position of *eve* seven stripes.

As an independent study of the dynamics of peak behavior we have computed the correlation of *x* positions of the peaks with the precise developmental age of embryos determined by measuring the degree of membrane invagination.
Unlike ANOVA, the correlation analysis can characterize the movement of expression domains continuously showing the tendency of movement with time. Table 3 shows the correlation results for *eve* gene. Higher values of Pearson correlation coefficient, *r*, correspond to the peaks which change their position with time, while the peaks which do not move have the lower correlation values. It is evident, that the results shown in this table coincide with those presented in Table 1. Coordinates of *eve* peaks plotted against the age of the embryo are shown in Figure 2.

Table 3. Correlation of positions of the *eve* peaks with the age of an embryo.

	1max	1min	2max	2min	3max	3min	4max	4min	5max	5min	6max	6min	7max
r	0.03	0.25	0.34	0.43	0.45	0.43	0.50	0.53	0.64	0.57	0.59	0.66	0.79
p<	0.74	0.01	10^{-3}	10^{-5}	10^{-5}	10^{-5}	10^{-8}	10^{-10}	10^{-12}	10^{-10}	10^{-10}	10^{-15}	10^{-15}

It is known that gap genes regulate pair-rule genes, and this further leads to the correct segmentation of *Drosophila* body. To understand whether shifts of gap and pair-rule gene expression domains are synchronous, we considered the correlations between changes in *x* positions of domain extrema in the embryos scanned for expression of both types of genes. The positions of maximal gap expression show high correlation with the movement of pair-rule peaks, which are located within their expression domains. Further we plan to analyze these data more thoroughly and to examine not only the positions of maximum expression of gap genes, but the positions of their domain borders as well.

Acknowledgements

This work is supported by NIH grants 4 RO3 TW01147-01 and RO1 RR07801-11 and GAP award RBO-1238.

References

1. Kosman D., Reinitz J., Sharp D. (1997) Automated assay of gene expression at cellular resolution. Pac. Symp. on Biocomput. 6-17.
2. Kozlov K., Myasnikova E., Samsonova M., Reinitz J., Kosman D. (2000). Method for spatial registration of the expression patterns of Drosophila segmentation genes using wavelets. Computational Technologies. 5, 112-119.
3. Merrill P., Sweeton D., Wieschaus E. Requirements for autosomal gene activity during precellular stages of Drosophila. Development. 104, 495–509 (1988).
4. Myasnikova E., Samsonova A., Kozlov K., Samsonova M., Reinitz J. (2001) Registration of the expression patterns of Drosophila segmentation genes by two independent methods. Bioinformatics. 17, 1, 3-12.
5. Nüsslein-Volhard C., Kluding H., Jurgens G**.** (1985). Genes affecting the segmental subdivision of the Drosophila embryo. Cold Spring Har. Symp.Quant. Biol. 50, 145-154.
6. Reinitz J., Sharp D. (1995). Mechanism of *eve* stripe formation. Mechanisms of Development. 49:133-158.
7. Reinitz J., Kosman D., Vanario-Alonso C., Sharp D. (1998). Stripe forming architecture of the gap gene system. Developmental Genetics. 23:11-27.

IIUDB: AN OBJECT-ORIENTED SYSTEM FOR MODELLING, INTEGRATION AND ANALYSIS OF GENE CONTROLLED METABOLIC NETWORKS

**** Freier A.*** [1]***, Hofestädt R.*** [1]***, Lange M.*** [2]

[1] Bioinformatics Workgroup, Faculty of Technology, Bielefeld University, Germany
[2] Plant Genome and Resource Center, IPK Gatersleben, Germany
e-mail: afreier@techfak.uni-bielefeld.de
*Corresponding author

Key words: *object-oriented modelling, pathway computation, gene networks*

Resume

Motivation: Our goal is the integrative construction and analysis of regulative gene networks based on the automatic access to available data sources. Retrieving the input data needed, the federative database approach provides online access to heterogeneous and physically distributed databases. In our group, we apply it to the integration of molecular data, covering the fields of genome, gene regulation, metabolism and disease. A system is needed to construct networks from integrated data and to prepare networks found. Supporting the analysis of different data types, the mapping from database to network structure should be configurable by the user. The further analysis of prepared networks can answer, e.g. the influence of gene regulation to metabolism or the reasoning of metabolic diseases.
Result: IIUDB is a toolbox to support the object-oriented modelling of gene network data, enabling the user to specifically integrate data from remote data sources. The integrated object networks are stored and used to assemble, analyse and visualize networks between biochemical objects, e.g. metabolic pathways. In the main part of our work we introduce a framework to derive pathways from integrated object networks directly.
Availability: The software has been written in Java and is available using Java Webstart under the URL: http://tunicata.techfak.uni-bielefeld.de/iiudb/servlet/iiudb.main.

Introduction

Today, more and more Internet databases are published providing selected molecular data concerning metabolism, gene networks and their application (Baxevanis, 2001). Actually, several systems for molecular database integration (Etzold et al., 1996; Stevens et al., 1999; Freier et al., 2002) are used get efficient access to distributed databases. At the same time information systems for modelling and visualization of regulative molecular networks are presented (Waugh, 2000; Goesmann et al., 2002; Glass, Gierl, 2002). But, even systems specialized at the same topic (e.g. gene networks) show differences by data modelling and differences by information content. On the one hand's side there is the evolution of molecular databases and on the other there are information systems, which have to be modified or rewritten to be able to be applied to new or changed content demandments, databases and topics. For object-oriented and distributed modelling of molecular data, CORBA (OMG, 1996) has been applied in the past (Coupaye, 1999; Kemp et al., 1999). Actually, existing systems implement object services providing a previously defined object structure, where application's methods exclusively are specialized to. Thus, a processing of user-specific objects is not possible. In this article, the system of *Individually Integrated User Databases (IIUDB)* is presented, which allows the runtime modelling of CORBA objects. The main idea of the system is to create user-defined object-oriented databases to store specific data found in public data sources in it before further processing them. While IIUDB integrates data into different user-defined databases, it at the same time enables the user to analyse and extract biochemical networks and pathways stored in every database. In our project *MARG (Modelling and Animation of Gene Regulative Networks)*, we apply IIUDB embedded in the *MARGBench* system.

The IIUDB System

The main task of IIUDB is the preparation of integrated gene network data to support network storage, analysis, simulation and visualization. The architecture contains three layers: Network Integration, Network Modelling and Network Analysis. At first, related data is stored using the object model. Here, we import data from external data sources and will at least find information about all objects known in gene networks. Then, a framework interconnects integrated objects by interaction objects in a directed graph as they participate at biochemical processes. Finally, pathways are computed and stored as prepared networks.
Modelling Object Networks: The initial step in our system's workflow starts with computational modelling of the real world objects. The user selects typical concepts, e.g. complex datatypes, inheritance and object references modelling the object type. For object specification of IIUDB databases, we use the *OMG IDL* language (OMG, 1996). Thus, a database scheme *S* is specified within an IDL document. It describes a network, if at least each object class refers to or is referred by

other object classes. In the domain of gene networks, we model object types, e.g. "Enzyme", "Pathway", "Gene" and others.

Creating Object Databases: The IDL class diagram will be processed by the tool *Database Builder*. Here, a database with CORBA interface capable to store objects of the users class diagram will be implemented and activated online. IIUDB handles multiple users and databases representing different user-specific views and content.

Integrate Data into Object Networks: The previously empty databases will be loaded with data from public molecular databases, representing the common domain knowledge. Here, available component data sources are queried, e.g. BRENDA for enzyme data, KEGG for metabolic pathways, RegulonDB for gene regulation, EMBL for gene annotation data and OMIM for disease information. To overcome the integration problem of heterogeneity and distribution, IIUDB cooperates with the *BioDataServer (BDS)* module (Freier et al., 2002), which provides a homogeneous database view to heterogeneous and distributed data sources. More precisely, for each data type in IIUDB the distribution to different data sources can be modelled. By that, every IIUDB database can be loaded online with data using database integration. The integration of objects leads to object networks. A standard object database system is used for database operations needed. The extension e of a class i holds all of its objects: $e_i = \{o_1, o_2, .., o_k\}$. The database extension E includes all class extensions. Together with the database scheme, a database D includes intension (scheme) and extension (objects): $D = (S, E)$.

Exploration of Object Networks: The database system used supports *OQL (Object Query Language)* (Cattel, Barry, 1997) database queries to retrieve objects specified by their attribute properties. Navigation through the object network is done by following object references. A navigation path t between two classes i_{begin}, i_{end} is defined by the class attributes followed up stepwisely: $t = (i_{begin}, \{a1, a2, ..., a_n\})$.

Object Interaction Framework: Towards the modelling of gene networks, the very natural representation of object-oriented formalisms seems to be suitable. In IIUDB, an object interaction framework has been implemented to analyse object networks. But, bioprocesses can be characterized as event objects referring to objects included in the event. The framework includes a formalism to describe the pattern p of a bioprocess and a mechanism to detect defined patterns in integrated databases. The main element of the framework is the event object v, specifying the topology of the bioprocess v = (IN, OUT, INF, LOC). Event objects v interconnect database objects in the way that they refer to input objects IN_v and output objects OUT_v. Objects influencing the event are referred to as influence objects INF_v. The physical location of the bioprocess is determined by location objects LOC_v. The principle applied to derive interaction event objects from the integrated object network should be explained briefly: we assume, that objects involved in one event are interconnected in the object network directly or transitively by path navigation. However, there must be access paths existing for *IN*, *OUT*, *INF* and *LOC*. An interaction model combines a set of event patterns with the events found: $M = (P, V)$.

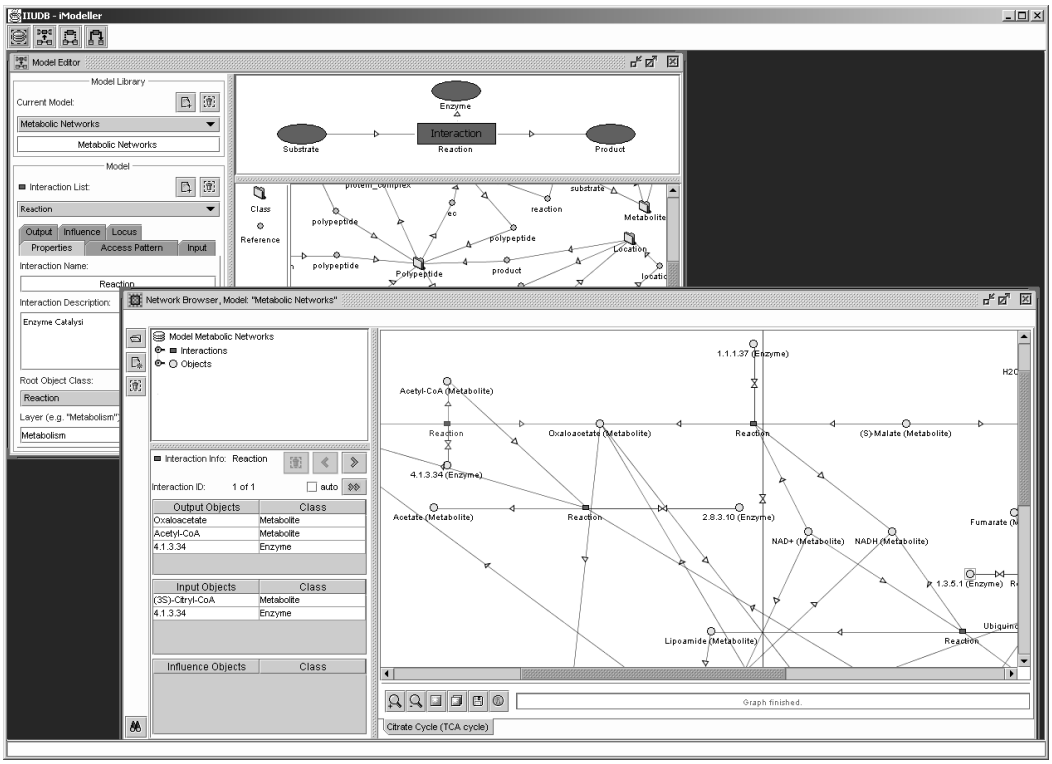


Fig. Ccomputing Citrate Cycle Topology from Integrated Object Networks.

Object Interaction Networks: So far, IIUDB databases contain the database scheme, the database extension and the interaction models: $D = (S, E, M)$. Event objects can be interconnected transitively as a pathway by the database objects they refer. Event objects of a pathway w are a subset of V: $w = \{v_1, v_2, ..., v_n\} : v_j \text{ in } V$. Starting at object o_1 in E of our network, the system knows about all consuming and producing events. Selecting an object o_2 as next checkpoint in our network, the system will insert all events leading to o_2 into our pathway w. Without any background knowledge, a search mechanism analysing the reachability of o_2 starting at o_1 can be used.

Conclusion

The growing number and the evolution of molecular databases demands for integrative and adaptive information systems. With IIUDB, we have been developing an object-oriented system supporting the preparation and analysis of gene controlled metabolic networks. While object-networks are constructed from data integration, a framework has been implemented, which is as simple as powerful to combine object networks with process modelling. The main result is the interactive and automatic assembling of bioprocess networks from integrated data. Actually, the system (see Figure) is available.

Acknowledgements

This work is sponsored by the German Research Council (DFG).

References

1. Baxevanis A.D. (2001) The Molecular Biology Database Collection: an update compilation of biological database resources. Nucl. Acid Res. 29, 1-10.
2. Cattell R., Barry D.K. (eds) (1997) The Object Database Standard: ODMG-93, Release 2.0. Morgan Kauffmann Publishers, San Francisco, CA.
3. Coupaye T. (1999) Wrapping SRS with CORBA: from textual data to distributed objects.. Bioinformatics. 15, 333-338.
4. Etzold T., Ulyanow A., Argos P. (1996) SRS: Information Retrieval System for Molecular Biology Data Banks. Methods in Enzymology. 266, 114-128.
5. Freier A., Hofestädt R., Lange M., Scholz U. (2002) BioDataServer: A SQL-based service for the online integration of life science data. In Silico Biology. 2.
6. Glass A., Gierl L. (2002) A system architecture for genomic data analysis. In Silico Biology, Special Issue: GCB'01.
7. Goesmann A., Meyer F., Kalinowski J., Giegerich R.(2002) PathFinder: reconstruction and dynamic visualization of metabolic pathways. Bioinformatics. 18, 124-129.
8. Kemp G.J.L., Robertson C.J., Gray P.M. (1999) Efficient access to biological databases using CORBA. CCP11 Newsletter, 3.1.
9. OMG (1996) The Common Object Request Broker Architecture: 2.0/IIOP Specification, OMG Document Number 96.08.04. OMG (Object Management Group).
10. Stevens R., Baker P., Bechofer S. (2000) TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. Bioinformatics. 16, 184-185.
11. Waugh M. (2000) Pathdb helps researchers analyze metabolism. Technical Report 1, National Center for Genome Resources.

MODELLING PLANT DEVELOPMENT WITH GENE REGULATION NETWORKS INCLUDING SIGNALING AND CELL DIVISION

*[1,2] Mjolsness E., [1,3] Jönsson H., [2] Shapiro B.E., [1] Meyerowitz E.M.

[1] Division of Biology, California Institute of Technology, Pasadena, CA 91125
[2] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109
[3] Department of Theoretical Physics, Lund University, Lund, Sweden
e-mail: emj@caltech.edu
*Corresponding author

Key words: *Arabidopsis, Shoot apical meristem (SAM), Cellerator, computer modelling*

Resume

Motivation: The use of computer models for the understanding of biomolecular systems is increasingly important. Multicellular models are especially useful in the context of developmental biological systems.
Results: We show how a simple model of gene regulatory networks combined with models of signaling and cell division can be used to simulate behavior of a real biological system, the shoot apical meristem in *Arabidopsis thaliana*.
Availability: http://www-aig.jpl.nasa.gov/public/mls/cellerator

Introduction

The shoot apical meristem of *Arabidopsis thaliana* is an example of a developmental system which can be modeled at genetic and mechanical levels provided that suitable mathematical and computational tools are available to represent intercellular signaling, cell cycling, mechanical stresses, and a changing topology of neighborhood relationships between compartments.
In previous work we have introduced a mathematical framework for gene regulation networks combined with cell signaling (Marnellos, Mjolsness, 1998), and the "Cellerator" package for automatic model generation from reactions relationships (Shapiro et al., 2000) and regulatory relationships along with cell division (Shapiro, Mjolsness, 2001). These tools may be combined to produce models capable simultaneously of transcriptional regulation, intercellular signaling, cell division, and mechanical deformation as appropriate to a developmental model. Here we apply this approach to developmental modelling to the case of the Arabidopsis shoot apical meristem (SAM).

Model

Generalizing from (Marnellos, Mjolsness, 1998) we use the combined gene regulation and cell-cell signaling dynamics:

$$\frac{d}{dt}v_a(t) = \frac{1}{\tau_a}\left[g(u_a + h_a) - \lambda_a v_a\right], \tag{1a}$$

where

$$u_a(t) = \sum_b T_{ab} v_b(t) + \sum_{I \in Nbrs} \Lambda^I \sum_b \hat{T}_{ab} v_b^I(t) + \sum_{I \in Nbrs} \Lambda^I \sum_b \sum_c \tilde{T}^{(1)}_{ac} \tilde{T}^{(2)}_{cb} v_c(t) v_b^I(t) \ . \tag{1b}$$

Here T is an intracellular gene regulation network, $\hat{T}$ is an intercellular network, and $\tilde{T}^{(1)}$ and $\tilde{T}^{(2)}$ represent a more detailed intercellular signaling network which separates the connection of receptors and ligands ($\tilde{T}^{(2)}$) from the connection of receptors and nuclear pathway target genes ($\tilde{T}^{(1)}$). To this is added a simple model for cell growth and cell division, which can be chosen from a variety of published models. The resulting system can now be simulated within Cellerator as otherwise described in (Shapiro, Mjolsness, 2001). Figure 1 shows a regular initial condition for a two-dimensional meristem simulation, including five cell types for expression domains and outer cell layers.

Results and Discussion

The dynamical behavior of a much simplified 2D model, containing only central zone (light gray) and rib meristem (dark gray) starting from a rectangular grid initial condition, is shown in Figure 2.

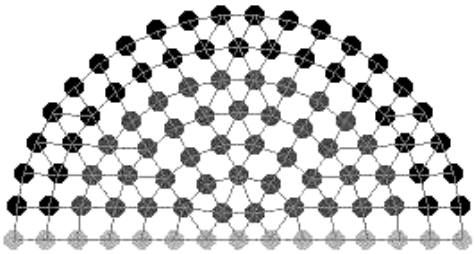


Fig. 1. Meristem initial conditions. Cell type is indicated by node shading.

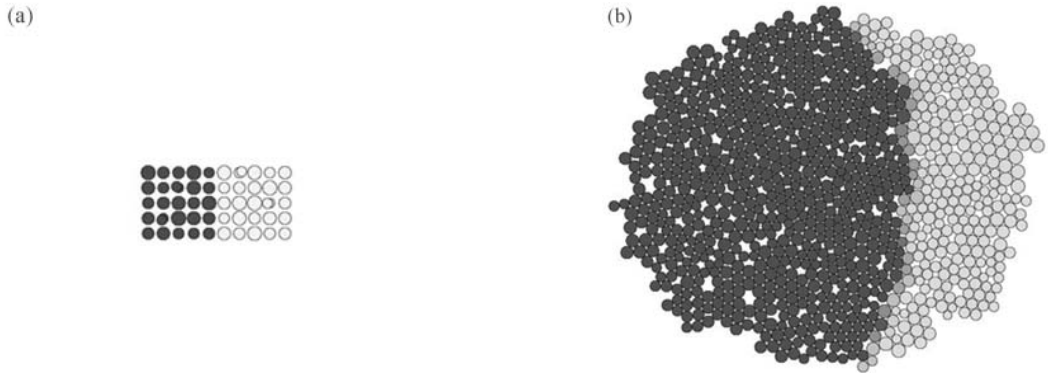


Fig. 2. Result of cell division, intercellular signaling, and intracellular gene regulation network dynamics. (a) Rectangular grid initial condition with equal numbers of cells in two expression domains. (b) Result of long-term dynamics.

In Figure 2, the spatial and growth dynamics is modeled by a reversibly breakable "spring" potential between neighboring cells as described in (Shapiro, Mjolsness, 2001). Each cell also have two proteins (PA, PB) and the protein concentrations follow the dynamics described in equation (1). Two cell types are defined by whether the concentration of PA is high or low indicated by colors (gray scale) in the Figure. Where PA is high, PB is low and vice versa (not shown).
The cells are initiated on a two dimensional grid with a small random deviation in size and growth rate (which determines the period of the cell cycle). There are two different initial protein concentrations of the cells, dividing them into two regions of cell types (PA concentration high/low).
In Figure 2a, the first cell division has just occurred and the cells are all in nearly the state in which they were initiated. Cells have just started to grow and move from their original positions. As time elapses further, the cells start to divide and by the time of Figure 2b the number of cells has greatly increased. Also the intercellular interaction leads to change in protein concentrations, converting cells of one type into the other (note medium gray cells in Figure 2b) when the two cell types are adjacent, as may be the case for the central zone and rib meristem of the SAM. This is also seen in Figure 2b by the asymmetry of the different cell type regions although they have the same division behavior. Note also the emergent hexagonal close-packing regions resulting from the nonlinear-spring elastic model described in (Shapiro, Mjolsness, 2001).
Future simulation work likely to be accessible with this model include the study of different rates of cell division in central zone and rib meristem, the migration of cells from rib meristem to stem, the dynamical stability of these three domains, the incorporation of the peripheral zone, and the lineage relationships within the separate layers of the shoot apical meristem.

Acknowledgements

The research described in this paper was carried out, in part, by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the U.S. National Aeronautics and Space Administration. Further support came from the Whittier Foundation, the ERATO Kitano Symbiotic Systems project, and the California Institute of Technology President's Fund. HJ was in part supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium.

References

1. Marnellos G., Mjolsness E. (1998) A Gene Network Approach to Modelling Early Neurogenesis in *Drosophila*. In Altman R.B., Dunker A.K., Hunter L., Klein T. (eds). Pacific Symposium on Biocomputing, World Scientific.
2. Shapiro B., Levchenko A., Mjolsness E. (2001) Automatic Model Generation for Signal Transduction with Applications to MAP-Kinase Pathways. In Kitano H. (ed), Foundations of Systems Biology, MIT Press, Cambridge, Massachusetts.
3. Shapiro B., Mjolsness E. (2001) Developmental Simulations with Cellerator. Second Intern. Conf. on Systems Biology (ICSB).

BIOUML – FRAMEWORK FOR VISUAL MODELLING AND SIMULATION OF BIOLOGICAL SYSTEMS

Kolpakov F.A.

Biosoft.Ru, Novosibirsk, Russia
Digital Design Technolgy Institute, SB RAS, Novosibirsk, Russia, e-mail: fedor@biosoft.ru

Key words: *systems biology, graphic notation, simulation, ODE, MATLAB, Java, GeneNet*

Resume

Motivation: With the completion of several genomics initiatives, including the Human Genome Project, researchers are poised to begin the next phase of elucidating how living systems function. Systems biology, a synergistic application of experiment, theory and modelling towards understanding biological processes as whole systems, requires integrated software environment that spans the comprehensive range of capabilities including access to databases with experimental data, tools for formalized description of biological systems structure and functioning, as well as tools for their visualization and simulations.
Results: Here we describe architecture and structure of BioUML framework designed for formalized graphic notation of biological systems structure and functioning, their simulations and access to databases on biological pathways. BioUML meta model provides an abstract layer to present structure of any biological system as a clustered graph. BioUML viewer and editor provide visualization of these graphs as diagrams and their editing. To incorporate any databases on biological pathways into BioUML framework we introduce a module concept and demonstrate it by creating module for GeneNet database. BioUML modeler allows a user to create and modify visual diagrams of biological systems and provides automatic generation of their executable models as MATLAB M-files. Using MATLAB these models can be simulated and investigated.
Availability: http://www.biouml.net; http://groups.yahoo.com/group/biouml.

Introduction

BioUML is designed as common purpose framework for systems biology providing formalized graphic notation of biological systems structure and functioning, their visualization and simulations as well as access to databases with relevant experimental data. BioUML is mostly oriented towards representing biochemical networks including cell signaling pathways, metabolic pathways, gene networks and molecular genetics systems.
BioUML framework is Java application consisting from following parts:

- *meta model* – provides an abstract layer to present structure of any biological system as a clustered graph.
- *BioUML viewer* – a universal viewer to visualize graphs of biological systems structure as diagrams.
- *BioUML editor* – universal diagram editor.
- *BioUML search engine* - it allows a user to create graphs of related biological entities. It provides similar functionality with TRANSPATH search engine (Schacherer et al., 2001); however the resulted graph can be edited and customized by a user using BioUML editor. Currently it is under construction.
- *BioUML modeler* - allows a user to model/simulate dynamics of biological systems using block diagrams.
- *Database modules* – provides incorporation of different databases on biological pathways into BioUML framework
- *Standard diagram and data types* – an attempt to standardize data types and graphic notations for biological pathways.

Meta model

The core of BioUML framework is meta model (Fig. 1) providing an abstract layer to present structure of any biological system as a clustered graph that further can be visualized as a diagram (Fig. 2) or stored as XML file (http://www.biouml.net/xml.shtml).
Class **DiagramElement defines** common attributes for all graph elements. Any instance of this class contains reference to corresponding object from a concrete database. By this way we wrap arbitrary database object to be element of diagram. To provide automatic executable model generations we can associate a role (variable or right side of equation) with any diagram element.
All graph edges are directed and are instances of `Edge` class. Simple graph nodes are instances of **Node** class. **Compartment** class is used to group several nodes in one compartment. **EquivalenceNodeGroup** is a special case of compartment to group nodes equivalent in a given context, for example homologous genes or proteins. All diagrams are instances of the same class **Diagram.** To take into account different diagram types, **Diagram** contains **type** attribute that is

instance of **DiagramType** class. **DiagramType** is responsible for visualization and semantic control of specific diagram. For this purpose it uses specialized subclasses of **DiagramViewBuilder** (to generate diagram view) and **SemanticController** to check the diagram semantics. **DiagramType** also defines what classes (database objects) can be used as nodes and edges.

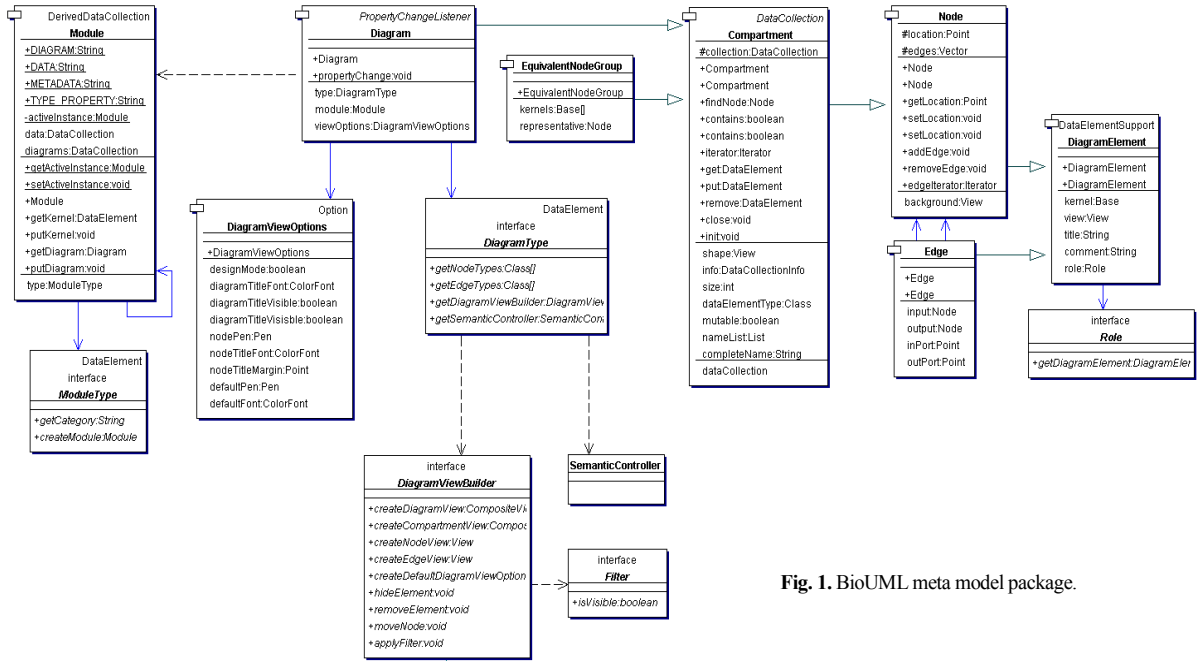


Fig. 1. BioUML meta model package.

Module concept

We introduce a concept of modules that serve for incorporation of different databases on biological pathways into BioUML framework. If we use a metaphor, then BioUML can be imagined as operating system (for example Windows), and modules are concrete programs (for example MS Word).

Concrete BioUML modules defines mapping of database content into diagram elements and diagram types that can be used with the database. As was described above, diagram type also specifies concrete classes of **DiagramViewBuilder** to generate database specific diagram view and **SemanticController** to check the diagram semantics.

Currently we have developed module for GeneNet database (Kolpakov et al., 1998) and now we are working on modules for TRANSPATH (Schacherer et al., 2001) and KEGG (Kanehisa et al., 2002) databases. We provide all sources of GeneNet module (http://www.biouml.net/genenet.shtml) to allow other developers to use them as an example to incorporate they databases into BioUML framework.

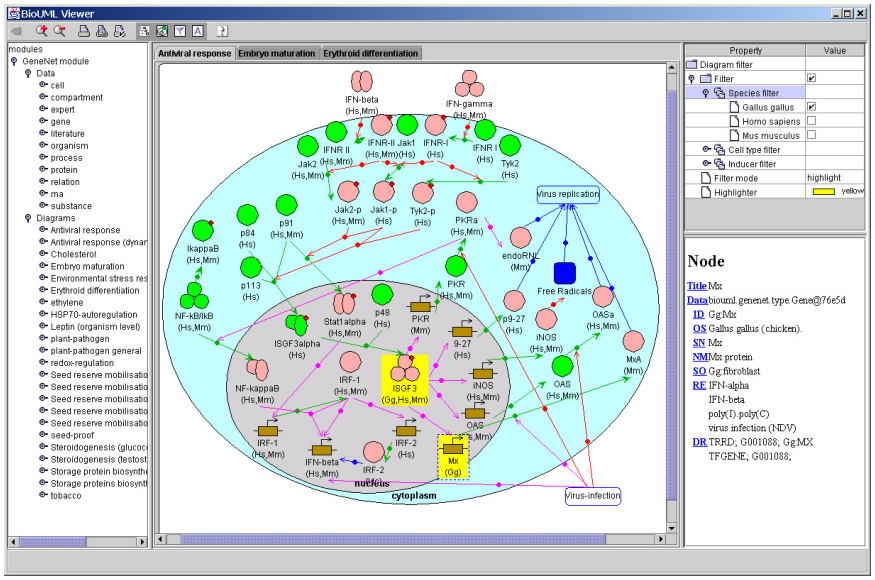


Fig. 2. BioUML viewer screen short. Left pane shows GeneNet database content that was incorporated as GeneNet module; central pane shows "Antiviral response" diagram; right top pane shows filters that are applied to the diagram, right bottom pane shows properties of "Mx" graph node. The data are taken from the publicly available version of GeneNet database (http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/).

Standard diagram and data types

We try to standardize data types and graphic notation for description biological pathways structure and they simulation using approach suggested in GeneNet system (Kolpakov et al., 1998) as a start point. Currently we specify data three diagram types (http://www.biouml.net/standard.shtml):
1) Pathway structure diagram – the diagram type to present metabolic and signal transduction pathways structure.
2) Pathway simulation diagram – extension of pathway structure diagram where variables are associated with graph nodes and differential equations – with graph edges.
3) Generalized pathway diagram – a pathway structure diagram generalized by different species, cell lines or experimental conditions.

BioUML modeler

BioUML modeler allows a user to model continuous dynamics systems that can be represented by system of ordinary differential equations (ODEs). A simulated biological system is presented as pathway simulation diagram (Fig. 3). To specify right side of differential equation MATLAB language is used, additionally some conventions is used for variable names. When such diagram is build, BioUML modeler allows user to automatically generate executable models as MATLAB M-files (Fig. 4) and start powerful MATLAB ODE suite (Shampine, Reichelt, 1997) for model simulations.
Example below demonstrates application of BioUML modeler for simulations simple pharmacokinetic model. Here 100 units of some drug A were injected intravenously. This drug can be break up by some enzyme E in liver giving the metabolite B. We suggest that drug flow from blood to liver is proportional to drag amount of in the blood. The same is true for drug flow from liver to blood, however the constant is other (k_1 in first case and k_2 in second case). We also assume that enzyme concentration in liver is E_0 and the dynamics of break up reaction can be described using Michaelis-Menten equation. Figure 4 shows the model diagram and result of its simulation, figure 5 demonstrates M-files that were automatically generated by BioUML modeler.
To simplify creation of complex diagram we are developing a set of standard reactions. It will be similar with standard blocks in MATLAB/Simulink – user should only specify some reaction constants while the needed differential equations will be generated automatically. The other direction of future current work is implementation of hybrid models to allow us joint modelling of a continuous subsystem with logical components.

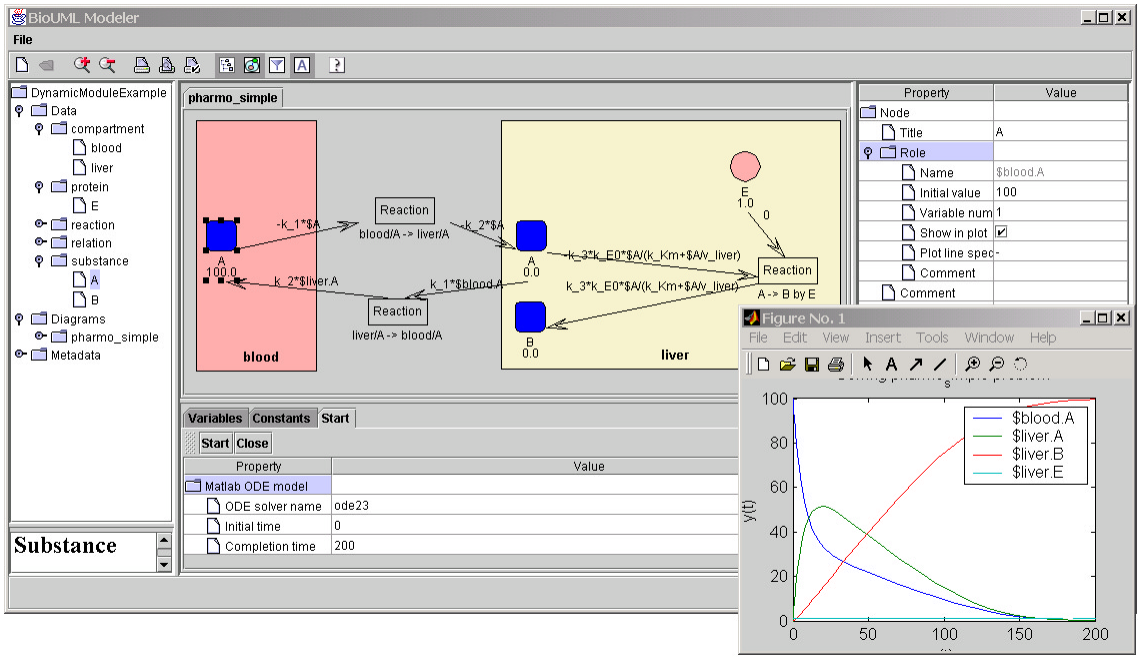


Fig. 3. BioUML modeler graphic user interface. Strings above arrows are right sides of differential equations associated with graph edges; numbers – are initial values of corresponding variables associated with graph nodes. Bottom right – result of model simulation using MATLAB.

```
%script for 'pharmo_simple' model simulation
%constants declaration
global k_1 k_2 k_3 k_E0 k_Km v_blood v_liver
k_1 = 0.1
k_2 = 0.05
k_3 = 0.01
k_E0 = 1.0
k_Km = 0.1
v_blood = 100.0
v_liver = 100.0

%Model variables and their initial values
y = []
y(1) = 100.0                          % y(1) - $blood.A
y(2) = 0.0                            % y(2) - $liver.A
y(3) = 0.0                            % y(3) - $liver.B
y(4) = 1.0                            % y(4) - $liver.E

%numeric equation solving
[t,y] = ode23('pharmo_simple_dy',[0 200],y)

%plot the solver output
plot(t, y(:,1),'-',t, y(:,2),'-',t, y(:,3),'-',t, y(:,4),'-')
title ('Solving pharmo_simple problem')
ylabel ('y(t)')
xlabel ('x(t)')
legend('$blood.A','$liver.A','$liver.B','$liver.E');

------------------------------------------------------------
function dy = pharmo_simple_dy(t, y)
% Calculates dy/dt for 'pharmo_simple' model.

%constants declaration
global k_1 k_2 k_3 k_E0 k_Km v_blood v_liver


% calculates dy/dt for 'pharmo_simple' model
dy = [ -k_1*y(1)+k_2*y(2)
       -k_3*k_E0*y(2)/(k_Km+y(2)/v_liver)-k_2*y(2)+k_1*y(1)
       k_3*k_E0*y(2)/(k_Km+y(2)/v_liver)
       0]
```

Fig. 4. Generated by BioUML modeler M-files to simulate 'pharmo_simple' model. Top - script file for model simulation and graphic result presentations; bottom - function to calculate dy/dt for the model.

Conclusion

We described the current status of BioUML framework. However our ultimate goal is to create of visual language for systems biology similar to UML. We would like to involve scientific community in this process and we provide special forum http://groups.yahoo.com/group/biouml/ for this purpose.

Acknowledgements

Part of this work was supported by the grant of Volkswagen-Stiftung (I/75941) and company DevelopmentOnTheEdge.com that provides its product BeanExplorer (www.beanexplorer.com) for development user interfaces for BioUML framework. Author is grateful to Sergey Zhatchenko and Alexander Kel for useful comments and discussions, as well as to Igor Tyazhev, Vlad Zhvaleev and Oleg Onegov for technical support.

References

1. Kanehisa M., Goto S., Kawashima S., Nakaya A. (2002) The KEGG databases at GenomeNet. Nucl. Acids Res. 30, 42-46.
2. 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(6), 529-537.
3. Schacherer F., Choi C., Gotze U., Krull M., Pistor S., Wingender E. (2001) The TRANSPATH signal transduction database: a knowledge base on signal transduction networks. Bioinformatics. 17(11), 1053-1057.

A SYSTEM FOR VISUAL MODELLING OF GENE NETWORKS' STRUCTURAL AND FUNCTIONAL ORGANIZATION

Loktev K.A. [1]***, Tkachev Yu.A.*** [2]***, Ananko E.A.*** [1]***, Podkolodny N.L.*** [1, 2*]

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
*Corresponding author: e-mail: pnl@bionet.nsc.ru

Key words: *gene networks, visual modelling, databases*

Resume

Motivation: In cells, tissues, organs and organisms, a huge amount of different processes that are controlled genetically take place simultaneously, including molecular genetical, biochemical, and physiological ones. For studying regulatory mechanisms of these processes and the ways of their preordered modification, it is necessary to develop effective computer methods for description, reconstruction, and modelling of complex gene networks.
Results: A system for visual modelling of structural and functional gene network's organization was developed. This system is compiled of: (1) a gene network editor realized in the environment Windows NT/2000/XP, (2) an application server, which provides logic operation in the system's functioning and connection to the database server, (3) a database server. Development of the application server and database server was made in the Oracle9i environment.

Introduction

Gene networks describing real physiological processes include hundreds and even thousands of components. Modelling of such complicate systems may represent a non-trivial process, which demands considerable efforts of many researchers.
The system presented in this work is aimed at visual modelling of the structural and functional organization of complex gene networks. The visual model may simplify the gene network description and analysis of its functioning. It could serve as a basis for developing a mathematical model of a gene network describing dynamics of gene networks.
Previously, we have developed the GeneNet system that laid the foundation for development of several dozens of gene networks, which describe regulation at the gene level of various functional processes in an organism (Ananko et al., 2002; Kolchanov, 2001). However, the experience of working with the previous version of the GeneNet system has revealed some limitations, when operating with complex gene networks with many elements. Additionally, the way of visualization was not perfect. In this connection, a necessity has appeared to develop the novel variant of the system with possibilities of its further extension, introduction of novel types of objects and setting up any type of graphical representation, together with using special approaches oriented to the complex gene networks.

Methods and Algorithms

The program system is based on the following main principles:
I) usage of the object-oriented visual modelling oriented to the complex gene networks;
II) possibility of an arbitrary extension of the object set, object attributes, and types of object visualization without making alterations in the program itself;
III) broad application of the XML/XSL technology.
The object-oriented modelling combines the process of the object-oriented decomposition, visual notation for description of logical, physical, statical, and dynamical models of the system described.
For each component of a gene network, we order its graphical representation, which could be modified by editing configuration files of a system. This enables to set up the system flexibly for different types of visual representations of gene networks.
In our program, we use the following types of decomposition of gene networks and approaches improving image sensing of complex systems:
I) In accordance with the levels of specification of a gene network's description. Following this approach, one may construct multi-level hierarchical models, when detailed composition of some system's component is hidden from a user at the level of description considered. Also, at each annotation stage, it is possible to enter incomplete information, when the mechanism of realization of some processes is unknown.
II) Spatial decomposition according to different types of compartments (cells, tissues, organs, etc.).
III) Temporal decomposition providing isolation of sub-systems according to typical periods, when the state of sub-systems alters.

IV) Usage of thematical layers (parts of a gene network) that are interpreted in terms of the subject area. The object set determining thematical layer is described by limitations on their attributes and ordered as the query to the database. Description of layers and the way of their composition are stored in the database.
V) «View» is a part of a gene network that determines the region of a user's interests at particular stage of the annotation process. A user clearly indicates, which object should be inserted or deleted from the field of vision in the course of a gene network description, which objects and relations will be used under inputting novel information and extension of a gene network. This possibility helps a user to concentrate an attention only on the actual at that moment fragments of a gene network. The objects located at the background are not active and pale-colored. When the work is continued, the last user's «view» on the gene network is conserved.
VI) The hierarchy tree of the elements of the scheme enables to operate by multi-component objects (they could be composed, copied, deleted, moved, and marked). This essentially lightens perception of hierarchical relations between the objects in complex systems.

Data Representation

Description of the main components of a gene network, i.e., proteins, RNAs, reactions, regulatory relations, etc. could be easily represented in an object-relational form. However, for gene network representation, it is necessary to use another type of the data model. A user's view on the data and results of the query to the database is a sort of a gene network, representation of which could be varying. This sort of data is referred to the weakly-structured types of data. For their representation in the present work, we apply the XML technologies.
Also, XML description is used for making queries for saving, modifying, and extracting information about the gene network's objects from the database, as well as for representing information related to operation of a system as a whole (restrictions for attribute values, storing of SQL texts of queries from server to database, configuration files of the system components, etc.).
To represent a gene network by the standard Web Browser, we have developed the program GeneNetViewer, which uses the SVG (Scalable Vector Graphics) format. This format enables to create qualitative zooming.
SVG drawings could be both dynamic and interactive. The object model of a document, Document Object Model (DOM) for SVG, enables to create effective animation based on application of the XML scripts. By such a technology, such events as, for example, mouse cursor moving over an object or mouse clicking could be handled.
For transforming information extracted from the database in the format needed, along with transformation of a gene network representation into SVG format and for accumulation of information in the database, we use the XSLT technology.

Implementation and Results

Three levels could be distinguished in the system architecture: (1) Oracle9i database; (2) application server Oracle9iAS, including as the components Container for J2EE (OC4J), XML SQL Utility, and XML Parser; (3) client programs that communicate with the server over the HTTP protocol, by which XML messages are sent. Currently, two client programs are realized: gene network editor and gene network viewer.
The database is realized within the frames of an object-relational Oracle 9i database by using technologies of the object types and nested tables. This approach enables to extract data stored in the database in the form of XML files. These data could be easily transformed (by XSL transformations) and send in a textual form over HTTP protocol. This database accumulates as systemic data related to functioning of the server and editor, as the data on biological objects. The data on biological objects have three levels of representation:
I) Abstract classes describing notions of the subject area or the object types (e.g., such notions as a protein, gene, RNA, etc.);
II) Specific realizations of different types of objects of abstract data (e.g., modified protein);
III) Description of the objects like essences, reactions, or administrations in the context of a particular gene network. It inherits the properties of classes of the 2nd level. However, some attributes at this level have the values that are defined only for a particular gene network.
The main functions of the server are the following:
I) supporting connection of several clients with the database server;
II) querying, designing, editing, and deleting of gene networks, objects, processes, controls, and relations in the database GeneNet, according to the client's query; administration of the database GeneNet;
III) support of security and authorization of access to the database;
IV) loading and saving of information in the database in the format XML, etc.
For each user, the category of access is determined, as well as the access permissions to particular resources. Realization of the server implements the possibility of extension, with the goal to add the module of gene network analysis and modelling. Interaction between the server and database is supported by technology JDBC and the query language SQL. The data obtained by the server are transformed into the XML format and sent to clients.
The program of the gene network visualization (GNViewer) is realized as the Java applet and oriented to usage of the standard Web Browser (like Internet Explorer 5.5 or higher, or Netscape) and Adobe SVG plug-in, which depict graphical

representation of a gene network obtained form the server as the SVG format. Usage of the XML parser in the Java applet enables to change visualization dynamically, on the client's side, to isolate the layers, and some other procedures improving visual analysis of the gene network.
The main functions of the gene network editor are the following:
I) displaying, input, and editing of gene networks represented in a form of a graph;
II) support of different types of decomposition of complex gene network;
III) constructing novel objects and editing of the attributes of already existing gene networks;
IV) writing of information about gene networks into the database;
V) loading of information about the gene network from the database;
VI) interaction with the server of the system by exchanging with XML messages over HTTP protocol;
VII) conveying of data to the data server for authorization of a user.
Editor of gene networks, GENED, is developed in the environment Windows NT/2000/XP in the language Visual C++. This editor uses the following additional components:
I) D2VectorEditor, an ActiveX of general purpose for manipulating by complex geometrical objects on the plain (author's elaboration);
II) MSXML parser, a COM object representing developed tools for operating with XML and XSL documents (Microsoft);
III) GDIPlus.dll, a library of the object-oriented two-dimensional graphics for Windows (Microsoft).

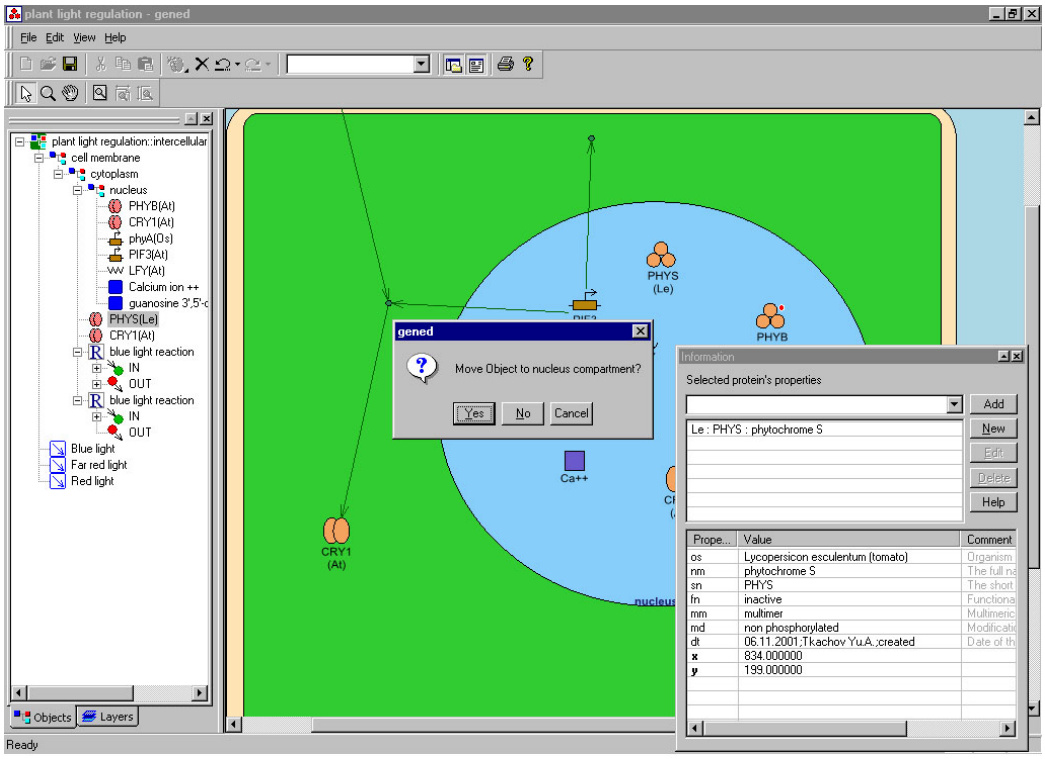


Fig. GENED screen short in the course of visual modelling of the structural and functional organization of a gene network.

At present, gene network editor has two basic ways (types) of visualization of a diagram: in a form of a tree (hierarchy of relations between the objects of the gene network) and in a form of two-dimensional image (gene network graph) (see Fig.). Although the types of visualization are fixed, a particular view of an image may be arbitrarily modified in dependence upon object attributes by editing configuration files.
To generic type of objects, with which the program operates, we refer, for example, such an object as compartment, that is, the object, which may contain the objects of the other types (compartments and simple objects). There is a possibility to extend the number of object types. All objects may have an arbitrarily extended set of properties that influence their visualization in the diagram. The set of properties (attributes), type of each property (either obligatory or not, type of a value; default value), admissible values of an attribute, etc. are ordered in special configuration files. The way how the object attribute influence on an object visualization, is also written in configuration files. To basic (generic) object properties, we refer, for example, coordinates on the plain (for all objects) and dimensions (for compartments).
All the data processed by the program, as well as all configuration files, are the XML documents, or XSL stylesheets.
Usage of the XML standard for data description produces flexible possibilities for adopting the editor to particular ways of gene network visualization. Object visualization may be constructed from the set of such graphical display elements as rectangle, oval, and line. Description of such visualization is made in the XML standard.

Conclusion

A system for visual modelling of structural and functional organization of complex gene networks was developed. The gene network editor may be flexibly edited for the pattern of visualization of the gene network elements. The editor enables to extend the number of types of objects, their attributes and limitations used for data verification.
The technologies applied for development the database and application server are capable to extend functional possibilities of the system and to operate with large heterogeneous data massives.

Acknowledgements

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

References

1. Ananko E.A., Podkolodny N.L., Ignatieva E.V., Podkolodnaya O.A., Stepanenko I.L., Kolchanov N.A. GeneNet system: its status in 2002. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002) (This issue).
2. Kolchanov N.A. Gene Networks Description and Modelling in the GeneNet System. In: Gene Regulation and Metabolism: Post-Genomic Computational Approaches eds. Collado-Vides J. and Hofestadt R. MIT Press (Book chapter), 2001.
3. Kanehisa M., Goto S., Kawashima S., Nakaya A. (2002) The KEGG databases at GenomeNet. Nucl. Acids Res. 30, 42-46.
4. Karp P.D., Riley M., Saier M., Paulsen I.T., Paley S.M., Pellegrini-Toole A. (2000) The EcoCyc and MetaCyc databases. Nucl. Acids Res. 28, 56-59.

A GRAPH-THEORETIC APPROACH TO COMPUTER ANALYSIS OF GENE NETWORK STRUCTURE

[*] ***Dobrynin A.A., Makarov L.I., [1] Podkolodny N.L.***

Sobolev Institute of Mathematics, SB RAS, Novosibirsk, Russia
[1] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
e-mail: dobr@math.nsc.ru, makarov@math.nsc.ru, pnl@bionet.nsc.ru
[*]Corresponding author

Key words: *gene network, computer analysis, graph algorithm, software*

Resume

Motivation: The modern biological databases accumulate plenty of information on genes and genetic processes. Efficient software is necessary for reconstruction and simulation of gene networks inspecting the functions of organisms.
Results: The basic software for manipulating with the structure of gene networks on the basis of graph-theoretic approach has been developed. These tools are intended for solving the problems as computing numeric and structural data for characterization of gene networks, searching for important domains and significant regulations in them, establishing similarity of gene networks, constructing a novel network from already existing ones, etc.

Introduction

A gene network is a molecular genetic system, in regulation of which the key role is played by the group of genes that are coordinated in functioning and interacting at realization of certain molecular, biochemical or physiological function of an organism (Kolchanov, 2001; Ananko, 2002). In connection with development of modern molecular genetic methods aimed at studying genome structure and gene expression mechanisms, as well as with accumulation of huge amount of factual material, the reconstruction of a gene network became available practically for each attribute of any organism: from viruses and bacteria to higher plants, animals, or humans.
Some approaches to computer representation of the gene networks have been developed at the Laboratory of Theoretical Genetics of IC&G SB RAS (Kolchanov, 2001; Ananko et al., 2002; Loktev et al., 2002). It has allowed describing dozens of gene networks supervising various processes in plant, animal or human organisms (see the database *GeneNet* at http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/).
The gene network describing the real physiological process includes hundreds and even thousand components (elementary objects, structures, molecular events, and processes). Revealing regularities in structure and function organization of a gene network of such complexity is possible only with the help of special computer programs. In this connection, one of the important problems is development of computer methods for the logic analysis of the structural and functional organization of the gene networks. Among the perspective approaches to resolving this problem is application of the graph theory methods (Harary, 1969). The methods of graph theory are known to be widely and effectively applied for studying the structure of organic molecules (King, 1983).
In this work, we present the basic software for logical analysis of the structure of gene networks on the basis on the graph-theoretic approach. These tools are intended for solving the following problems: computing numeric and structural data for characterizing gene networks, finding its important domains and significant regulations, establishing the similarity of gene networks, constructing a new network from the others, etc.

Methods and Algorithms

1) Data representation
Basic constituents of a gene network are molecular-genetic objects: genes, RNAs, proteins and protein complexes, low-molecular non-proteinaceous substances, organs, tissues, cells, cellular compartments, etc., as well as the interactions between these molecular-genetic objects. The interactions are classified into two classes:
1) reaction, a molecular event forming a novel object (e.g., transcription, translation, multimerization, maturation of proteins, etc.) and
2) regulatory relation, an event that switches on, enhances, suppresses or completely switches off a process.
Notice that initial representation of a gene network in the GeneNet format makes difficult to apply directly the graph theory methods to gene networks analysis. Preliminary, a gene network should be represented as a formal mathematical construction of graph theory. There exist several ways how to describe a gene network as a formal graph-theoretic object. To this aim, it is convenient to use representation of a gene network as a labeled directed graph. Thus, the vertices of the graph correspond to the objects, as well as to the object interactions. A vertex label has the hierarchical structure. Attributes of the first level are *entity, regulatory event, reaction, process*, etc. A vertex-entity may be any of the following:

gene, protein, RNA, substance, etc. The set of attributes for a vertex-regulation includes *increase, decrease, switch on, switch off, direct* or *indirect*.

In this case, a gene network can be interpreted as a special type of semantic network, with interpretation of vertices as entities (genes, RNA, proteins and protein complexes, low-molecular non-proteinaceous substances, organs, tissues, cells, cellular compartments, etc.) and interactions (reactions, regulatory events, etc.). Graph arcs reflect semantic relationships (for example, participation in a process) and roles that biological entities play in a molecular-genetic process (substrate, product, enzyme, etc.).

The initial information for analyzing gene networks was extracted from the *GeneNet* database, which accumulates description of the structural and functional organization of more than 20 gene networks. The structure and data of these networks can be examined at http://wwwmgs.bionet.nsc.ru/mgs/gnw/genenet/. The representation of a gene network described in *GeneNet* was converted to the format of vertex adjacency list. For each vertex, all its neighbor vertices are listed.

2) Search for a strongly connected subnetworks in the gene network graph

The number of pathways that provide regulatory mechanisms in a cell determine the complexity of gene networks. These pathways may vary in length and interact with each other in a complex way (simple regulations, positive or negative feedback pathways). To improve our understanding of the regulatory principles, we should detect such regions for further analysis. Therefore, we are interested in constructing the maximal number of subnetworks, in which every vertex (vertex-entity) can be accessible from any other vertex. Such a subnetwork corresponds to a strongly connected component in the respective graph. The developed program searches for all strongly connected components on the bases of the algorithm described elsewhere (Reingold et al., 1977). As an example, let us consider the gene network on *Erythroid differentiation* (Podkolodnaya et al., 2000). It has two strongly connected components illustrated in Fig. 1.

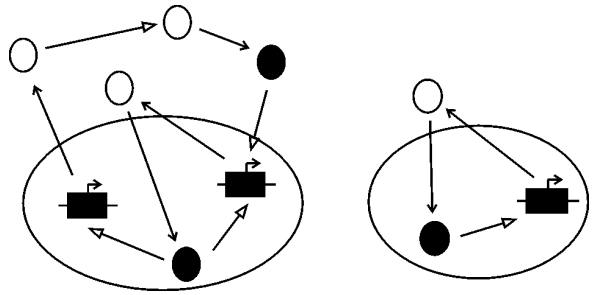


Fig. 1. Strongly connected network regions.

3) Search for circuits of gene network graph

In gene networks, some pathways perform regulatory functions, thus, generating the positive and negative feedbacks in a network. Usually, the negative feedback supports the equilibrium in a cell. For instance, in case cellular concentration of a substance becomes too high for a normal functioning, a regulatory mechanism switches on and decreases the quantity of this substance until the concentration falls to a norm. Then the regulatory mechanism switches off, and the process is repeated by oscillating pattern. In order to detect a regulation, we should analyze the circuits of a gene network graph. It is clear that strongly connected components contain all circuits. A type of a pathway is mainly defined by the label of vertex-relations (increase or decrease) of circuits. A program for generating all possible circuits has been developed. For example, the Cholesterol biosynthesis regulation network has two rather long pathways that are potential candidates for further detailed analysis (see Fig. 2).

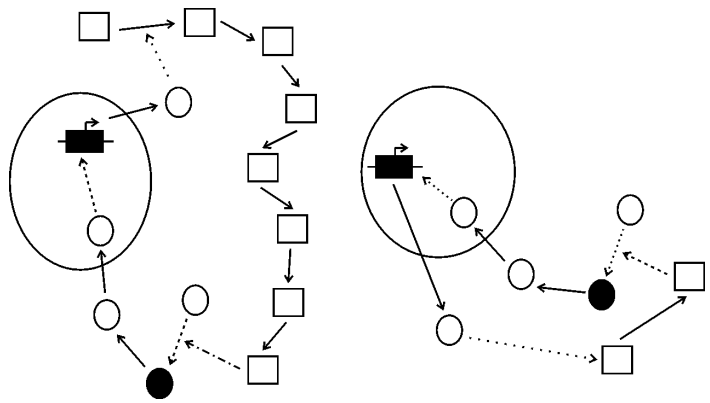


Fig. 2. Two regulatory pathways in *Cholesterol biosynthesis regulation* network.

4) Search for critical gene network elements (graph cutpoints)
A random mutation is capable to destroy objects in a gene network. Various regions of a network are characterized by different vulnerability to damaging. After removing a critical object, more disconnected parts could appear in a gene network, i.e., some objects lose communication between them. Therefore, it is desirable to locate such weak places. Graph cut points correspond to such critical elements. To increase network robustness, we can insert additional elements into a network. The elaborated computer program finds various connectivity characteristics of a graph network.
5) Basic operations under networks
The efficiency of gene network analysis significantly depends on our ability to extract a network desired from databases or to construct a novel network from selected ones. The computer program allows us to make basic operations under networks. Among them are union, subtraction, intersection, etc. Here by intersection we mean "set-theoretic" ope-ration (but not searching for maximal common parts between networks). Another examples of operations are: reconstruction the maximal subnetwork, with a given set of network objects (genes and proteins); or finding the gene regulatory region within the region where the factor acts.
6) Calculation of gene network microstructural parameters
For a gene network graph, various numeric and microstructural parameters may be calculated. Thus, there are many graph invariants that can by applied for characterizing the structure of networks. The Wiener index is an example of such a parameter (Dob-rynin et al., 2001). It is equal to the sum of distances between all vertices of a graph. The mean distance is the average distance between vertices. The data presented below contains some parameters and statistical data for *Erythroid differentiation* network for all organisms (the distance invariants are considered for the largest connected component having 169 vertices):
number of graph vertices: 215
number of graph arcs: 235
average degree of vertices: 1.1
maximal vertex outdegree: 18
maximal vertex indegree: 4
number of strongly connected components: 2
number of nontrivial blocks: 10
number of connected components: 13
number of cutpoints (belong to nontrivial blocks): 8
number of circuits: 3
number of circuits in which all vertices-relation have the attribute *increase*: 3
number of circuits having a vertex-relation with the attribute *decrease*: 0
network mean distance: 10
network diameter: 25
network radius: 13.
7) Search for all embeddings of a given fragment into network
In some cases, the function of a gene network is processed via operation of the same parts in both networks. The created computer program finds all embeddings of a given fragment into a network. This option can be also applied for aggregation (or expansion) of gene networks by contracting the unessential or typical fragments.
8) Search for classes of networks with similar structures
Suppose we have a set of networks (graphs) and we need to separate them into classes such that the networks within each class possess by a similar structure. To represent a network, we should select its suitable structural descriptors. For instance, we can fix a list of network fragments of some kind (say, pathways with specified proteins and reactions). For every network, we count the number of entries of such fragments and assume that the vector obtained describes the network. These vectors form a vector space of points, for which we may calculate mutual distances. Then for this distance matrix we apply the taxonomy comparison method. Vectors-descriptors are given often as a collection of different invariants. As a result one can obtain: classes of networks with similar structures, typical elements in these groups, classification of novel networks. The taxonomy program is based on the algorithm described in (Makarov, 1998).

Conclusion

The basic software for manipulating with the structure of gene networks and logical analysis of its structurally functional organization has been developed on the basis of graph-theoretic approach. These tools are intended for solving the following problems: computing numeric and structural data for characterizing of gene networks, finding its important domains and significant regulations, establishing the similarity of gene networks, constructing a new network from other ones, etc.
The elaborated software is a part of the system for computer analysis of gene networks developed at the Laboratory of Theoretical Genetics of IC&G SB RAS.

Acknowledgements

This work was financially supported in part by the Russian Foundation for Basic Research (grants № 01-07-90376 and 02-07-90359), Russian Ministry of Industry, Sciences and Technologies (grant № 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Projects № 65). The authors are grateful to N.A.Kolchanov and E.A.Ananko for helpful discussions.

References

1. Kolchanov N.A. (2001) Gene networks description and modelling in the GeneNet system. In: Collado-Vides J., Hofestadt R. (eds). Gene Regulation and Metabolism: Post-genomic Computational Approaches. MIT Press.
2. Ananko E.A., Podkolodny N.L., Ignatieva E.V., Podkolodnaya O.A., Stepa-nenko I.L., Kolchanov N.A. (2002) GeneNet system: its status in 2002. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
3. Loktev K.A., Tkachev Yu.A., Ananko E.A., Podkolodny N.L. (2002) A system for visual modelling of gene networks' structural and functional organization. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002) (this issue).
4. Podkolodnaya O.A., Stepanenko I.L., Ananko E.A., Vorobiev D.G. (2000) Representation of information on erythroid gene expression regulation in the GENEEXPRESS system. Proc. II Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS' 2000), 1, 34–36.
5. Harary F. (1969) Graph Theory. Addison-Wesley, Reading, MA.
6. Reingold E.M., Nievergelt J., Deo N. (1977) Combinatorial Algorithms. Theory nd practice. Prentice-Hall, Englewood Cliffs.
7. Dobrynin A.A., Entringer R., Gutman I. (2001) Wiener index for trees: theory and applications. Acta Appl. Math. 66, 211–249.
8. Makarov L.I. (1998) Methods and algorithms for the prediction of chemical compound properties by common fragments of molecular graphs. J. Struct. Chem. 39, 93–102.
9. King R.B. (ed). (1983) Chemical Application of Topology and Graph Theory. Elsevier, Amsterdam.

COMPUTER SYSTEMIC BIOLOGY: INFORMATIONAL AND SOFTWARE TOOLS FOR COMPLEX MOLECULAR BIOLOGICAL SYSTEMS

[1] Kolchanov N.A., [12] Podkolodny N.L., [1]Likhoshvai V.A.*

[1] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia, e-mail: kol@bionet.nsc.ru
[2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
[*]Corresponding author

Key words: *systemic biology, gene networks, mathematical simulation*

Resume

Motivation: Events occurring on molecular genetic level initiate the actions resulting in a systemic integrative biological effect at the levels of cell, tissue, organ, or entire organism. Research into the mechanisms of such effects with the aim to search for efficient means allowing these process to be controlled requires approaches realized in specialized computer systems that combine methods of bioinformatics, datamining and knowledge discovery, mathematical and computer simulation of gene networks, and optimal control.
Results: Architecture of an integrated system for studying complex molecular genetic systems was developed. A number of its components have been realized.

Introduction

Research into mechanisms underlying molecular interactions depending on the genetic information and specific features of molecular structures may shed the light on biochemical functions and roles of elementary components as well as on the specific control patterns of gene networks. These pieces of knowledge form the background for computer simulation of gene networks allowing changes in molecular genetic, biochemical, physiological, morphological, and other characteristics of various organisms to be predicted as well as optimal control actions and stimuli for correcting genetically specified impairments of the body operation to be searched for.
For this purpose, a new generation computer technologies integrated in the computer system **GeneNetDiscovery** is developed at the Siberian Branch of the Russian Academy of Sciences. This system provides solving a wide range of problems in the field of computer analysis and simulation of complex molecular genetic systems (gene networks, genetically controlled metabolic pathways, signal transduction pathways, etc.) including (i) accumulation of data and knowledge on the structure–function organization of gene networks; (ii) integration of the information on gene networks and metabolic pathways; (iii) construction of gene network mathematical models and their computer-assisted numerical analysis; (iv) study of dynamic behavior of complex molecular genetic systems (gene networks) in norm, in case of pathologies and metabolic diseases, and under the effect of adverse environmental factors at molecular genetic, cellular, and organismal levels; and (v) search for optimal control of gene networks and correction of their behavior in the case of various pathological states.

Methods and algorithms

The system GeneNetDiscovery is a tool designed for creating gene network models that could be further used in another operation system and for a variety of purposes.
For annotators and modelers, specialized worksites are developed in a Windows NT/2000/XP environment. Web interfaces are developed for outside "casual" users. Currently, two client programs are realized (gene network editor and gene network viewer), which communicate with the server over HTTP protocol, by which XML messages are sent. Oracle9i is appled for controlling the databases. When developing the system's middleware controlling the logics of its operation and its linkage to the knowledge bases and databases, application server Oracle9iAS is used; it includes Container for J2EE (OC4J), XML SQL Utility, and XML Parser as components. The operation logics of the system GeneNetDiscovery is shown in Fig. 1.

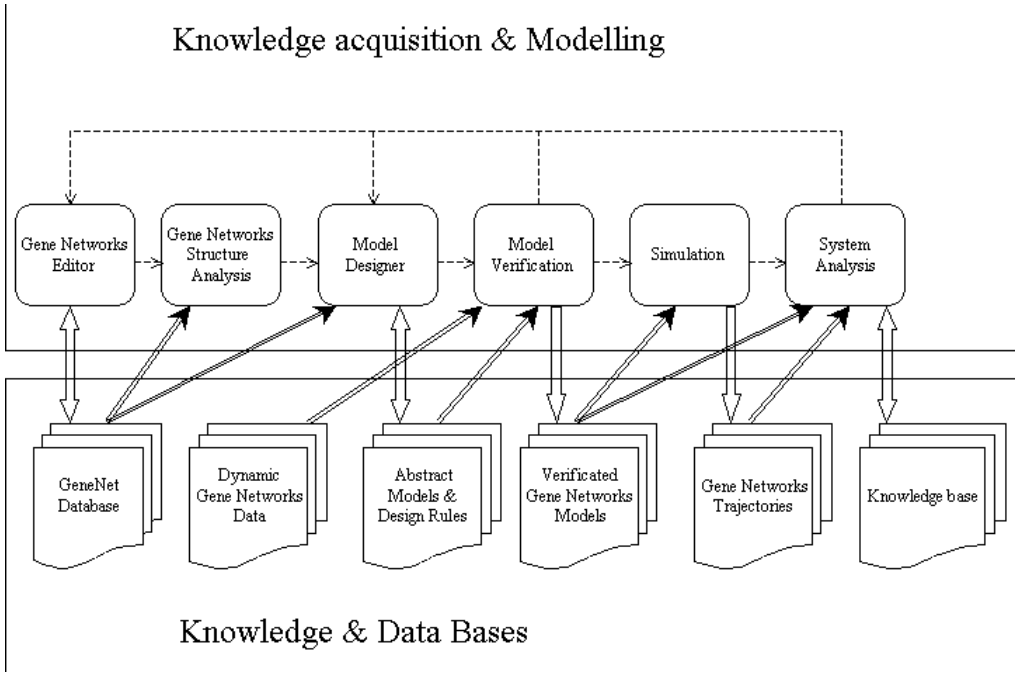


Fig. 1. Functional layout of the system GeneNetDiscovery.

The system GeneNetDiscovery comprises the following functional modules:
1)Subsystem for constructing models, including
GeneNet database, which accumulates the experimentally confirmed information on the structure–function organization of gene networks; dynamical data, etc. (Ananko *et al.,* 2002); a new version of this database in an Oracle9i environment is under development now (Loktev *et al.,* 2002);
Module GENED (gene network editor) for visual simulation of the structure–function organization of complex gene networks and data input into GeneNet (Loktev *et al.,* 2002); and
Module for interactive construction of gene network models using the information accumulated in the GeneNet database (as systems of differential equations, logic automatons, or hybrid models describing gene network dynamics);
2) Subsystem for analyzing gene network models, including
Module for computer logical analysis of the structure–function organization of gene network basing on a graph-theoretic approach (Dobrynin *et al.,* 2002) and
Module GeneNetStep for analyzing numerically gene network models and studying qualitatively the behavior of solutions of differential equations in the space of model's parameters (Fadeev *et al.,* 2002);
3) Subsystem for identifying models, including
Module for solving inverse problems, that is, identifying models and assessing their parameters from experimental data, expert estimations, and experimentally observed trajectories of system behaviors (Likhoshvai *et al.,* 2002); and
Database of dynamical data describing behavior of molecular genetic subsystems under various experimental conditions (Likhoshvai *et al.,* 2002);
4) Subsystem for simulating gene networks and analyzing their behavior, including
Module for calculating trajectories of behavior of molecular genetic systems and accumulating the calculated variants in database;
Module for computer analysis of gene network dynamics basing on data mining and knowledge discovery approach; search for gene network behavior patterns, their analysis, and generalization for creating the knowledge base with the system (for testing of certain algorithms see Borisova *et al.,* 2002);
Knowledge base compiling variants of the calculated trajectories of gene network behaviors and pieces of knowledge used for constructing models; and
Application of the models developed for solving various scientific problems, in particular, studying the effect of mutations on gene network dynamics and molecular mechanisms underlying various pathologies stemming from impairments of the gene network operation (Ananko *et al.,* 2002; Stepanenko and Grigor'ev, 2002; Turnaev and Podkolodnaya, 2001; Nedosekina and Ananko, 2002; Kudryavtseva and Stepanenko, 2002; and other); and
5) Subsystem for controlling gene networks, including
Module for analysis of mutational portraits of gene networks and detection of optimal targets for their pharmacological regulation (Ratushny *et al.,* 2002) and

Module searching for optimal control of gene network operation in norm and pathology, in particular, for correction of organismal pathological states with account of their individual genotype-specific features (Latypov *et al.,* 2001a; 2002b)

Conclusion

Described in the work is the architecture of the system that is developed basing on multilevel computational approaches combining genome-encoded information with nongenomic network connections. A part of the modules of the system GeneNetDiscovery has been realized. The major algorithms forming the core of this system have been tested by solving particular problems.

Acknowledgements

This work was funded in part by the Russian Foundation for Basic Research (grants Nos. 01-07-90376 and 02-07-90359); Russian Ministry of Industry, Science, and Technologies (grant No. 43.073.1.1.1501); Siberian Branch of the Russian Academy of Sciences (Integration Project No. 65); US National Institutes of Health (grant No. 2 R01-HG-01539-04A2); and US Department of Energy (grant No. 535228 CFDA 81.049).

References

1. Ananko, E.A., Podkolodny, N.L., Ignatieva, E.V., Podkolodnaya, O.A., Stepanenko, I.L., and Kolchanov, N.A. (2002). GeneNet system: its status in 2002. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
2. Dobrynin, A.A., Makarov, L.I., and Podkolodny, N.L (2002). A graph-theoretic approach to computer analysis of gene network structure. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
3. Loktev, K.A., Tkachev, Yu.A., Ananko, E.A., and Podkolodny N.L. (2002). Система визуального моделирования структурно-функциональной организации генных сетей. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
4. Likhoshvai, V.A., Latypov, A.F., Nedosekina, E. A., Ratushny, A.V., and Podkolodny, N.L. (2002). Technology of using experimental data for verification of models of gene network operation dynamics. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
5. Borisova, I.A., Zagoruiko, N.G., Likhoshvai, V.A., Ratushny, A.V., and Kolchanov, N.A. (2002). Диагностика мутаций на основе анализа динамики генных сетей. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
6. Fadeev, S.I., Berezin, A.Yu., Gainova, I.A., Kogai, V.V., Ratushny, A.V., and Likhoshvai, V.A. (2002). Разработка программных средств в области математического моделирования динамики генных сетей. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
7. Kudryavtseva, A.N. and Stepanenko, I.L. (2002). Gene network of glutathione homeostasis: a response to oxidation stress. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
8. Latypov, A.F., Nikulichev, Yu.V., Likhoshvai, V.A., Ratushny, A.V., Matushkin, Yu.G., and Kolchanov N.A. (2002a). A method of solving problems of optimal control in dynamics of gene networks. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
9. Latypov, A.F., Nikulichev, Yu.V., Likhoshvai, V.A., Ratushny, A.V., Matushkin, Yu.G., and Kolchanov N.A. (2002b). Problems of control of gene networks in a space of stable states. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
10. Nedosekina, E.A. and Ananko, E.A. (2002). Генная сеть активации макрофагов при действии интерферона-гамма и липополисахаридов. Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
11. Ratushny, A.V. and Likhoshvai, V.A. (2002). Computer analysis of the effects of mutations in LDL receptor gene on the regulation of cholesterol biosynthesis in the cell. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
12. Ratushny, A.V., Likhoshvai, V.A., and Kolchanov, N.A. (2002). Analysis of mutational portraits of gene networks. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*
13. Turnaev, I.I. and Podkolodnaya, O.A. (2002). Генная сеть контроля клеточного цикла. Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
14. Stepanenko, I.L. and Grigor'ev, S.A. (2002). Organization of the gene network of apoptosis. *Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002).*

STRUCTURAL STABILITY OF *DROSOPHILA* CONTROL GENE SUBNETWORKS: COMPUTER EXPERIMENTS, QUANTITATIVE AND QUALITATIVE EVALUATION

**** Galimzyanov A.V., Tchuraev R.N.***

Institute of Biology, Ufa Research Center, Russian Academy of Sciences, Ufa, Russia, e-mail: galim@anrb.ru
*Corresponding author

Key words: *control gene networks, Drosophila melanogaster, Arabidopsis thaliana, λ-phage, structural and parametric stability, computer experiments*

Resume

Motivation: When investigating by means of mathematical methods the dynamics of gene networks controlling ontogenetic processes it is necessary to evaluate the structural, and particularly parametric, stability of constructed models. Alongside with independent interest, the solution of this problem allows to verify the adequacy of the models to some extent, since according to V.I.Arnold (Arnold, 2000) only "mild" models are able to correspond to real objects.
Results: With computer experiments on the models of three gene expression control systems it is shown that (1) gene networks controlling the development of organisms display a high degree of parametric stability; (2) DNA sites, being specific to regulatory proteins, permit synonymic substitutions in regard with ontogenetic processes; (3) possible neutralization of mutative phenotypical manifestations is particularly due to highly-organized structure of the control cellular gene network.

Introduction

A system is said to be structurally stable if it is stable to fluctuations in parameters and functions (Arnold, 2000). In the case of control gene networks (CGN) we understand the structural stability as a gene network ability to preserve its inherent normal regimes of functioning (stationary, transitional or periodic) under changes of the network structure in one of the following manners: either by addition/removal of a gene (genes), informational link (links) or their combination, and also under changes of kinetic parameters. The action thresholds of regulatory proteins and their complexes depend mainly on the degree of affinity to the proper sites, which is determined in primary nucleotid sequences of the proper sites, as are unity intensities of gene transcription, mRNA translation, transcript and protein degradation and parameters of RNA processing and transportation. Thus, the fluctuations in the values of kinetic parameters reflect the changes in the primary structure of genome DNA molecules. Let us suppose that the sensitivity evaluation of function regimes in the model to random fluctuations of the parameters in a relatively wide range of values is the evaluation of both parametric and structural stability of the model.
Previously an approach was elaborated to test the parametric stability of prokaryotic and eukaryotic molecular-genetic systems of gene expression control, namely the system controlling the λ-phage development (Ratner, Tchuraev, 1978) and the subsystem controlling *Arabidopsis thaliana* flower morphogenesis (Tchuraev, Galimzyanov, 2001). The present work formulates and solves the problem on the evaluation of parametric stability of the gene subnetwork controlling the early ontogenesis of *Drosophila melanogaster* fruit fly (Dr-CGN), the model of which (Tchuraev, Galimzyanov, 2001) was constructed on the basis of the method of generalized threshold models (Tchuraev, 1991) in its program realization (Galimzyanov, 2000).

Methods and computer experiments scheme

In the genetic block model (Tchuraev, 1991) describing the mechanism of mRNA and protein synthesis as well as the logic of interactions of regulatory substances and DNA sites we take into account the following kinetic parameters: a_{1j} and a_{2j} are the unity intensities of gene transcription and mRNA translation; b_{1j} and b_{2j} are the transcript and polypeptid degradation coefficients; P_{ij} are the regulatory protein threshold concentrations, etc. To study the model's parametric stability we performed computer experiments, in which the model functioned at the sets of values in the kinetic parameters chosen randomly either in wider intervals or within the range of *l*-percent (*l*= 10%, 20%, 30%, 40%) deviation from the parameter's values of a "good" set (when the model functioned in the normal regime). The gene network dynamics in each of one hundred nuclei in the control line (Reinitz, Sharp, 1995) along the anterior-posterior axis (with account for its location with respect to the gradient of morphogene concentrations) at a chosen random set of parameters was compared with the gene network dynamics of the same nucleus in the normal regime on the basis of the set of genetic blocks being in the active state. The stability of the Dr-CGN model to the fluctuations in the values of kinetic parameters was evaluated by the average number of normal regimes of the gene network functioning in all the nuclei of the total number of random

choices. *The parametric stability coefficient* C_{PS} was determined by the following expression: $C_{PS}=\frac{1}{n}\sum_{i=1}^{n} c_i$, where n is the length of the control line (n=100 nuclei), c_i is the percent of normal regimes of gene network function in the i-th nucleus of the total number of random choices (parametric stability local coefficient).

Implementation and Results

For the purpose of investigation an original PST-lab program module has been developed that realizes computer experimental scheme on testing the parametric stability of *Drosophila* gene subnetwork arbitrary fragments under the fluctuations in the values of kinetic parameters of an arbitrary subset. This module is an extension of the AGENDY computer program package (Galimzyanov, 2000) designed for modelling the dynamics of prokaryiotic and eukaryotic gene networks on the basis of the method of generalized threshold models. With the mg-fragment of the model consisting of maternal and gap genes the computer made 500 random choices for each type of the experiments. The table gives coefficients C_{PS}, gained in each experiment. The figure presents an example of a curve for evaluating the parametric stability of the model's mg-fragment to variations in the values of kinetic parameters in one of the experiments.

Table. Parametric stability coefficients of the model.

No.	Coefficient C_{PS} (in %) at different *l*				
	-	*l*=10%	*l*=20%	*l*=30%	*l*=40%
1	73.2	-	-	-	-
2	-	97.8	93	88.6	83.5
3	-	76.2	59.8	51	45.3
4	-	74.2	57	47.5	41.1

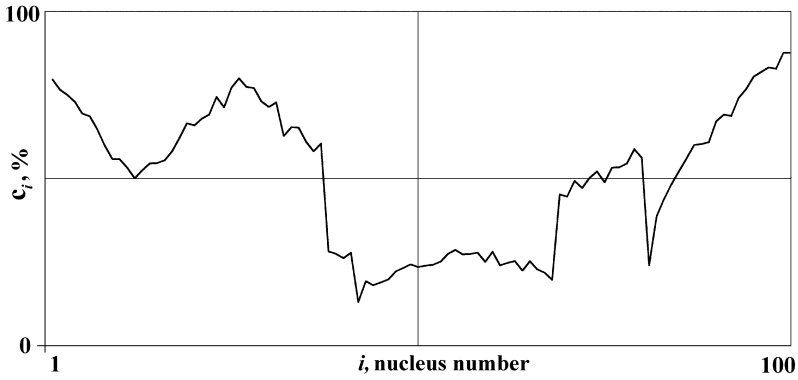


Fig. 1. The curve for evaluating the parametric stability of the model's mg-fragment in nuclei in the control line. The abscissa, numbers of nuclei (anterior pole is to the left, posterior pole is to the right); the ordinate, parametric stability local coefficient.

The experimental results show that (1) fluctuations of the thresholds decrease the model's parametric stability much stronger as compared to fluctuations of the RNA and protein synthesis and degradation unity intensities; (2) the model's parametric stability reduces with an increase in the proportion of deviations of the kinetic parameters values from a "good" set; (3) *Drosophila* embryo is heterogeneous in terms of the gene network parametric stability in the nuclei in the control line; the lowest parametric stability is found at the compartment boundaries, what might be explained by alternative functioning of genes in the subnetwork; (4) the less is the difference between maximum and minimum concentrations of some morphogene at the compartment boundaries (gradient local amplitude), the higher is the model's sensitivity to fluctuations in the threshold values of morphogene concentrations; (5) neutralization of mutative phenotypical manifestations may occur at the expense of the *Drosophila* gene network structure.

Discussion

The discussed control gene network consisting of maternal, gap, pair rule, segment polarity and homeotic genes is a hierarchic system with some elements of heterarchy associated with the presence of feedback loops. Genes of the lower levels are affected by genes of the higher levels and also of their own level, that provides genes' cascade-type functioning in time, when, for example, homeotic genes are activated later than those of segment polarity. As follows from the cascade-type Dr-CGN structure, the higher is the level of the changeable elements, the more negative is the influence of these fluctuations upon the normal function of all the system as a whole. Together with Bcd, Nos and Cad morphogenes, gap genes form the Dr-CGN upper (basal) level; hence, the hypothesis seems to be realistic assuming *high* and *uniform* parametric stability of the gene network mg-fragment under consideration in the nuclei in the control line. Some of our experimental data, however, do not agree with this assumption. Thus, it may be supposed that there are some other mechanisms not accounted for in the model that provide the formation of gene expression normal patterns in the *Drosophila* early ontogenesis. On the other hand, *critical stages* are known to exist during embryogenesis characterized, in particular, by higher sensitivity of cells and tissues to the effect of various external and internal factors and correspondingly by the lower CGN parametric stability. In the *Drosophila* early ontogenesis a specific stage of the

embryonal structural formation corresponds to the functioning of each class of genes mentioned above. In this aspect the obtained results should be treated as a direct corroboration of the fact that the syncytial blastoderm stage of the *Drosophila* embryo is the most vulnerable from the viewpoint of parametric stability.

References

1. Arnold V.I. (2000). Rigorous and mild mathematical models. Moscow Center of Continuous Mathematical Education, Moscow. Russian.
2. Galimzyanov A.V. (2000) Software automated package for analyzing the dynamics of control gene networks. Proc. of the BGRS'2000, Novosibirsk. 1, 233-234.
3. Ratner V.A., Tchuraev R.N. (1978) Simplest genetic systems controlling ontogenesis: organization principle and models of their function. In Rosen R., Snell F.M. (eds.). Progress in Theoretical Biology. 5, 81–127.
4. Reinitz J., Sharp D.H. (1995) Mechanism of *eve* stripe formation. Mech. Dev. 49, 133-158.
5. 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.
6. Tchuraev R.N., Galimzyanov A.V. (2001) Modelling of actual eukaryotic control gene subnetworks with the method of generalized threshold models as the base. Mol. Biol. (Mosk.). 35, 6, 1088-1094.
7. Tchuraev R.N., Galimzyanov A.V. (2001) The solution of the problems on parametric stability for ontogenesis control gene networks. Information and simulation systems for the analysis of gene regulation and metabolic pathways. Dagstuhl Seminar 01261, Report 313, 28-30. http://www.dagstuhl.de/DATA/Seminars/01/

TECHNOLOGY OF USING EXPERIMENTAL DATA FOR VERIFICATION OF MODELS OF GENE NETWORK OPERATION DYNAMICS

Likhoshvai V.A., Latypov A.F. [*1], Nedosekina E.A., Ratushny A.V., Podkolodny N.L. [2]

Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[1] Institute of Theoretical and Applied Mechanics, SB RAS, Novosibirsk, Russia, e-mail: latypov@itam.nsc.ru
[2] Institute of Computational Mathematics and Mathematical Geophysics, SB RAS, Novosibirsk, Russia
[*]Corresponding author

Key words: *gene networks, mathematical simulation, dynamic databases, adaptation of models, scenario technology*

Resume

Motivation: Development of adequate methods for analyzing the gene network operation laws involves elaboration of a technology for verification of parameters of gene network models on the basis of dynamic data and also methods of representation of experimental data in biological databases.
Results: A technology is proposed for identification of parameters of mathematical models of gene networks on the basis of scenarios reproducing the test protocol and test conditions. Two approaches are suggested: 1) evolutionary method and 2) method of successive iterative solution of the problem of model identification on the basis of a series of experiments.

Introduction

The main distinctive feature of the currently existing biological databases on the gene structure, metabolic pathways, and enzymatic processes is the fact that the main data accumulated there are the qualitative descriptive information and also the quantitative information of the static nature: constants of enzymatic reactions, protein masses, and lengths of nucleotide chains.
At the same time, the information on dynamics of behavior of biological systems is scarce. Meanwhile, information of this kind is necessary for solving many problems, including the development of adequate mathematical models of functioning of biological systems.
The greater part of kinetic data is currently scattered in thousands of scientific papers, and this field of knowledge is rather poorly structured, which complicates the use of these data.
The main feature of these data is their heterogeneity. The reason is that each set of dynamic experimental data is obtained under different test conditions, by different experimental actions, and at different times, i.e., test protocols and test conditions are significantly different for different sets of data. Ignoring this fact, one cannot adequately use the entire ensemble of experimental data to analyze the model. In addition, it is often necessary to involve indirect information on the behavior of the system examined or its subsystems. In particular, a typical problem is the necessity of using data obtained in experiments with different species.
To solve these problems, one needs new technologies that ensure systematic accumulation of dynamic characteristics of the behavior of biological systems in specialized databases and the use of the latter for identification of models.
A format for representation of dynamic experimental data on concentrations of compounds and substances with allowance for operation conditions of the object examined and types of experimental actions is proposed in the present paper, a dynamic database is described, and a technology for identification of parameters of mathematical models of gene networks on the basis of scenarios reproducing the test protocol and conditions is suggested.
This technology is tested by an example of investigation of particular molecular genetic systems: system for cholesterol synthesis regulation and system for macrophage activation.

Implementation and Results

The approach to identification of a mathematical model on the basis of heterogeneous sets of experimental data is based on the idea of implementation of an extended model that reproduces the protocol, all the registered conditions of each particular experiment, and all the types of experimental actions on the system under study. Thus, conditions of particular experiments are reproduced, and then the mathematical model is nested into these conditions as into an ambient medium. We called this nesting a ***scenario.*** Thus, one scenario reproduces the conditions of one (rarely more) experiment.
It is assumed that each experiment whose results are used to adapt model parameters describes the functioning of some subsystem or its small fragment.
Obviously, the use of a large body of inhomogeneous data requires the construction and simultaneous calculation of a large number of scenarios.
Each experiment is assigned an estimate of the degree of adequacy of the experimental situation to the system examined or a weight of deviations of dynamic variables of the model from experimental data in the overall quality functional.

These estimates are initially prescribed by an expert. In the course of model fitting, the contribution of each type of the experiment to the overall quality functional of the model is evaluated, and then the estimates are corrected.
In the general case, there are many local minimums of the quality functional of the model, which may have a specific biological interpretation and may refer to specific self-consistent classes of experimental situations.
The technology developed is based on solving three separate subtasks. The first subtask is the development of methods of annotation and extraction of dynamic data from scientific publications and their accumulation in databases. The second subtask involves the development of methods of reproduction of test conditions in the form of scenarios and methods of nesting gene network models into these conditions. The third subtask is the development of methods for analyzing mathematical models nested into a set of scenarios.

Subtask 1. Description of dynamic experimental data
Dynamic data are described in the format of the GeneNet database (Kolchanov et al., 2000; Ananko et al., 2002). For this purpose, the database format is supplemented with new informational fields. They allow one to introduce all the necessary data on compounds and substances into the database in a unified format: general characteristics (type of the compound, brief and full names of the substance, organism); information on the origin of the compound (cellular type, tissue, organ); data on the stages of development of the organism, stages of differentiation of the cell or cellular cycle; data on the state of the organism (whether it is healthy or has some pathology); data on the sex of the organism, data on the method of compound detection, and also information on cell concentration if the experiments are performed on dedicated cells or cellular lines. For substances used as external factors acting on a system (cell, organ, tissue, organism, etc.), the basic characteristics of the action are described: the source of obtaining these compounds, concentration, duration of the action, and additional characteristics of the action (temperature, acidity, etc.).
To enter information about the compound being measured, the database has special fields for the measurement time, concentration of the substance, and estimate of measurement errors in the form of standard deviations.

Subtask 2. Technology of constructing scenarios
Such a technology was implemented within the framework of the generalized chemical kinetic method of modelling (Likhoshvai et al., 2001). Modelling of scenarios, as well as modelling of gene networks, is performed in terms of elementary processes. The notion of the process is extended and can be used to describe all actions that can be performed in the course of an experiment. For instance, pulsed addition/deletion of a particular substance at the time T is simulated using a discrete logical block, which traces the current time in the system and replaces, at the time T, the "current" concentration of the substance by the value implied by the scenario used. More complicated patterns are also simulated in the form of a sequence of elementary processes of different nature. The general basis of elementary processes makes it possible to reproduce scenarios of arbitrary complexity. Codes for model calculations are supplemented with algorithms that allow simultaneous calculation of several scenarios. This is the basis for the development of flexible technologies for solving specific problems of mathematical simulation.

Subtask 3. Verification of parameters of gene network models
To seek optimal parameters of the model nested into S scenarios, we used two approaches described below.

Evolutionary method
Optimal parameters are chosen by the method of imitation of evolution of a population of individuals. It is assumed that each individual of the population is a carrier of the modeled system with an individual set of parameters, which are generated randomly. Each evolutionary act consists of two steps. At the first step, the individuals of the population reproduce. Each individual gives birth to a certain number of descendants. Reproduction is accompanied by random changes in parameters; as a result, an extended population consisting of maternal and mutant filial individuals is formed. At the second step, a fixed number of the most fitted individuals is chosen. The method is implemented as a computer program. The prescribed parameters of the model are the number of the chosen individuals, the number of descendants reproduced by one individual, and the model of mutation generation (several alternative models can be used). The fitness is treated in terms of the agreement of results obtained using the model and available experimental data.

Iterative method
Let the measured characteristics $X^e(t)$ of an object be related at each time moment t to the initial data X_0, parameters $a(t)$ characterizing the test conditions, and uncontrolled quantities $\xi(t)$ by the functional relation

$$X^e(t) = X^e(X_0, a(t), \xi(t), t). \quad (1)$$

Let the mathematical model be defined by the algorithm

$$X(t) = X(X_0, a(t), p(t), t), \quad (2)$$

where $p(t)$ are parameters of the mathematical model to be determined.

We assume that the measurements are performed at discrete times t_i, $i = 0,1,...,N$, with uniformly distributed errors, so that

$$X_i^e = X^e(t_i) + \delta X_i, \;\; a_i = a(t_i) + \delta a_i \; . \quad (3)$$

Then the parameters of the mathematical model can be determined by solving the following optimization problem using the principle of the minimum root-mean-square deviation:
Problem I-1.

$$F = \frac{1}{2}\sum_{j=1}^{J} \sigma_j \sum_{i=1}^{N} \left(X_i^j - \overline{X}_i^{je}\right)^2 \Rightarrow \min_{p\in R_m,\lambda,\mu},$$
$$X_0 = \overline{X}_0 + \lambda\cdot\delta X,\ \lambda\in[-1,+1],\ \lambda=\{\lambda_1,...,\lambda_J\}, \quad (4)$$
$$a = \overline{a} + \mu\cdot\delta a,\quad \mu\in[-1,+1],\ \mu=\{\mu_1,...,\mu_K\}.$$
$$X(t) = X(X_0, a(t), p(t), t), t\in\{t_i, i=\overline{1,N}\}$$

The bar refers to the measured values, m is the number of identified parameters, J is the number of measured characteristics, σ_j are the weight coefficients, and K is the number of parameters a.

If the inequalities

$$\left|X_i^j - \overline{X}_i^{je}\right| \le \delta X_i^j, \quad (5)$$

are satisfied in solving problem I-1, the mathematical model is adequate, otherwise, it is necessary either to change the model or to restrict the area of its applicability. Let inequalities (2) be satisfied. Then the following problems are solved to estimate the limiting values of the identified parameters.

Problem I-2a. $\Phi = \frac{1}{2}\sum_{m=1}^{M} p_m^2 \Rightarrow \min_{p\in R_m,\lambda,\mu}$

Problem I-2b. $\Phi = \frac{1}{2}\sum_{m=1}^{M} p_m^2 \Rightarrow \max_{p\in R_m,\lambda,\mu}$.

Both problems are solved under conditions (3) and (5).
Denoting the solutions as p' and p'', respectively, we obtain the values of parameters and the error estimate

$$\overline{p} = \frac{1}{2}(p' + p'') + \delta p,\ \delta p = \frac{1}{2}|p'' - p'|. \quad (6)$$

We assume that $p'' \ge p'$, otherwise, a permutation is necessary. It is important that these estimates determine ***the minimum range*** of admissible values of the identified parameter.

Now let we have a set of experimental scenarios $s = \overline{1,S}$ (Eq. (1) being valid for each of them) and algorithms of the type (2). We indicate the characteristics and internal parameters of the *s*th scenario by the superscript "s". We solve problems I-1, I-2a, and I-2b successively for each scenario. As a result, we obtain the upper and lower values of internal parameters p'^s and p''^s. We single out the kth parameter (the lowercase k here is a subscript, in contrast to K) common for S_k scenarios and find

$$P'_{k\max} = \max P_k'^j,\ P''_{k\min} = \min P_k''^j,\ j\in S_k.$$

For $P'_{k\max} \le P''_{k\min}$, the value of the kth parameter and the error are determined similar to Eq. (6), otherwise, we successively eliminate the numbers of scenarios where the maximum $P'_{k\max}$ and minimum $\mathrm{P}''_{\mathrm{k\,min}}$ are reached. The eliminated scenarios should be further subjected to additional consideration. Since the functions $p_k = g(\Phi)$ are nonmonotonic in the general case, we have to solve problems I-1 for eliminated scenarios with fixed domains of definition for identified parameters common with other scenarios. If inequalities (5) are satisfied in the course of solving these problems, the experiments and mathematical model are adequate; otherwise, it is necessary to consider the possibility of changing the mathematical model and/or revise the experimental data.

Conclusions

A technology is proposed for identification of parameters of mathematical models of gene networks on the basis of scenarios reproducing the test protocol and test conditions. The efficiency of the approach is tested by an example of constructing models of particular molecular-genetic systems: system for cholesterol synthesis regulation (Ratushny et al., 2002) and system for macrophage activation (Nedosekina et al., 2002).

Acknowledgements

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

References

1. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002) GeneNet: a database on structure and functional organization of gene networks. Nucl. Acids Res. 30, 398-401.
2. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaia O.A., Ignat'eva E.V., Goriachkovskaya T.N., Stepanenko E.L. (2000) Gene networks. Mol. Biol. (Mosk). 34, 533-544.
3. Likhoshvai,V.A., Matushkin,Yu.G., Ratushny,A.V., Anan'ko,E.A., Ignat'eva,E.V., Podkolodnaia,O.A. (2001) A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk). 35, 1072-1079.
4. Nedosekina E.A., Ananko E.A., Likhoshvai V.A. (2002) Construction of mathematical model of the gene network on macrophage activation under the action of IFN-g and LPS. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
5. Ratushny A.V., Likhoshvai V.A. (2002) Computer analysis of the influence of mutations in the LNP receptor gene on the system of regulation of cholesterol biosynthesis in a cell. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

COMPUTER ANALYSIS OF THE EFFECTS OF MUTATIONS IN LDL RECEPTOR GENE ON THE REGULATION OF CHOLESTEROL BIOSYNTHESIS IN THE CELL

Ratushny A.V. *, Likhoshvai V.A.

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

Key words: *computer analysis, gene network, cholesterol biosynthesis, mutation, LDL receptor*

Resume

Motivation: Study of the normal and pathological cholesterol transports from blood plasma into the cell mediated by LDL receptors has become a topical problem. Impairments of this system play a key role in development of hypocholesterolemia and atherosclerosis of humans and various animals.
Results: The mathematical model simulating cholesterol biosynthesis in the cell and its exchange with blood plasma cholesterol developed earlier with additionally verified values of its parameters was used for computer analysis of the effects of various mutations in LDL receptor gene on the system in question.
Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/

Introduction

The main fraction of cholesterol in blood plasma is contained in low-density lipoproteins, or LDL (Murray et al., 1988). Cholesterol is transported into the cell via a system mediated by LDL receptors. This system maintains a certain LDL level and, correspondingly, a certain cholesterol level, in blood plasma. The relevant experimental data demonstrate that impairments of this system play a key role in development of hypocholesterolemia and artherosclerosis of humans and various animals (Klimov, Nikul'cheva, 1999).
So far, over 300 mutations in LDL receptor gene have been detected and described. The major part of these mutations are large deletions or rearrangements, while the rest result from deletions of one or several base pairs or, rarer, insertions, or point nucleotide substitutions (Hobbs et al., 1990; Soutar, 1992).
In this work, computer analysis of the effects of various mutations in LDL receptor gene on the system studied was perfromed.

Methods and Algorithms

A computer model of gene network functional dynamics developed earlier (Ratushny et al., 2000) was used for the analysis. The model was developed in the context of chemical kinetic approach to simulation and is a system of ordinary differential equations describing the operation of this gene network supplemented with a set of discrete expressions used to imitate the external factors (Likhoshvai et al., 2001). Additional information is available from (Ratushny et al., 2000) and the papers used to describe this gene network in GeneNet (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/).
Identification of the parameters of mathematical model. The values of reaction constants (the parameters in question) were determined differently. All the parameters used in this mathematical model fall into the following groups: (1) constants of enzymatic reactions; (2) constants of macroprocesses (transcription, translation, etc.); and (3) the constants characterizing interactions of the gene network regulating cholesterol biosynthesis with its cellular and organismal environment. Values of the majority of enzymatic constants are compiled in the electronic databases WIT (http://www-unix.mcs.anl.gov/compbio/), BRENDA (http://brenda.bc.uni-koeln.de/), and other.
Values of the constants of macroprocesses were selected taking into account the biological data on their typical rates. For example, the translation rate constant is taken equal to 0.1 sec^{-1}, as (i) the distance between A-sites of the neighboring ribosomes in mRNA cannot be less than 20–60 nucleotides due to steric limitations and (ii) the mean elongation rate amounts to about 3–10 codons/sec (Spirin, 1986).
The parameters characterizing interactions of gene network with environment were estimated from general biological grounds, such as lifespans or half-lives of gene network components, their equilibrium concentrations, contents per cell, durations or rates of the processes considered, etc.
The values of parameters absent in the literature were verified using evolutionary method (Likhoshvai et al., 2002). Fig. 1 exemplifies the achieved compliance of the model with experimental data.

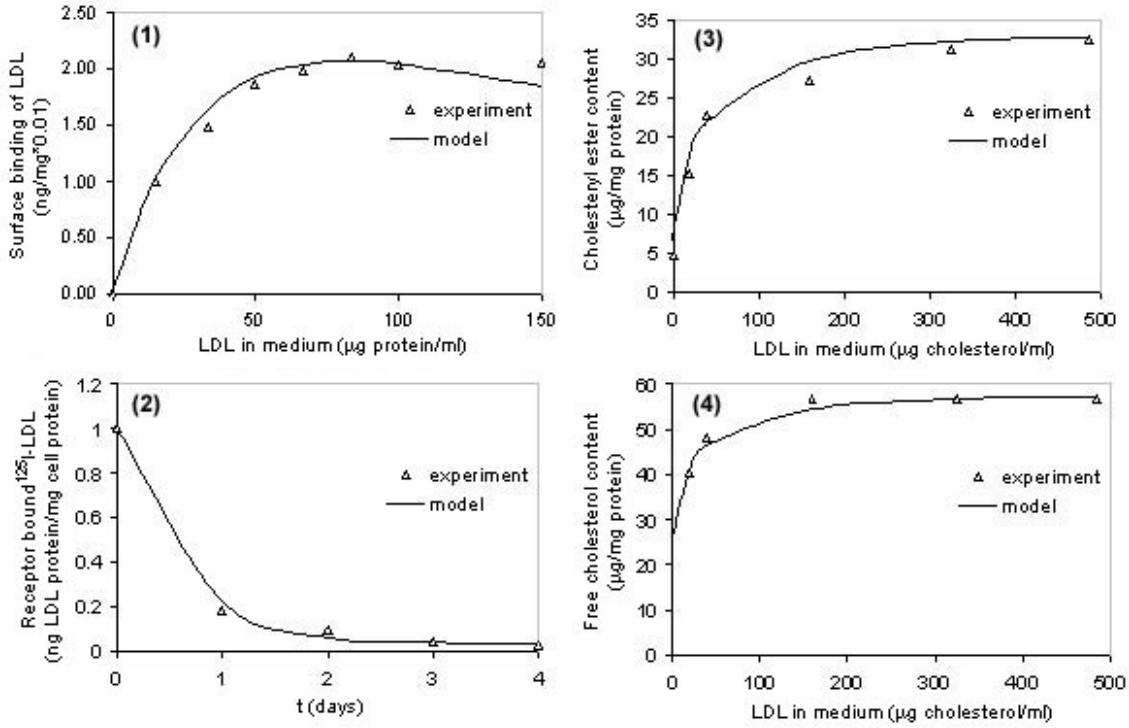


Fig. 1. Comparison of the calculations made using the model with experimental data: triangles, experimental data from (**1**) Brown & Goldstein (1979), (**2**) Goldstein *et al.* (1977), and (**3, 4**) Goldstein *et al.* (1975); full lines, simulations.

Implementation and Results

The model allows stationary characteristics and dynamics of the gene network, both in norm and in the presence of mutations, upon various effects to be studied.

Bold lines in Fig. 2 demonstrate the calculated response of the studied gene network in norm to a twofold increase in the inflow of LDL into blood plasma continuing over 8 h (hatched region). These conditions cause a monotonic increase in blood LDL, reaching an approximately fourfold level by 10 h of the experiment (**a,** n). In this process, the concentration of free receptors on the cell surface decreases (**c,** n), whereas the concentration of cholesterol in the cell changes insufficiently (**b,** n), which is explained by the negative feedback decreasing the rate of cholesterol biosynthesis in the cell upon its increased inflow from outside the cell. All the variables of this system take stationary values approximately 6 h after the internal effect is stopped.

Fine lines in Fig. 2 show the behavior patterns of the system when three different mutations are introduced. The first mutation (m_1; Fig. 2) exemplifies the class of mutations in LDL receptor gene preventing formation of immunodetectable protein (the so-called null alleles). Specific of a considerable fraction of the null alleles is increased concentrations of LDL receptor mRNA in cells of patients (Klimov, Nikul'cheva, 1999). Our numerical calculations simulating a twofold decrease in the rate of LDL receptor synthesis in the gene network demonstrate that the stationary number of free receptors on the cell surface decreases approximately 2.5-fold (m_1, **c;** Fig. 2). The stationary LDL concentration in blood increases approximately 1.5-fold, while the free cholesterol content in the cell decreases by 15% (m_1; Fig. 2). We explain a relatively small decrease in the intracellular cholesterol content with the ability of cell to compensate for the decrease in the external cholesterol inflow with its increased intracellular synthesis in combination with the negative feedback regulation of the cholesterol biosynthesis rate.

The graphs m_2 in Fig. 2 exemplify the class of mutations retaining the normal synthesis of LDL receptor and its transport to the cell surface; however, it displays a decreased ability to bind LDL. Most frequently, these mutations result from the deletion removing repeats 1 and 2 from the ligand-binding domain of LDL receptor gene (Russel et al., 1989). Using the model, we studied the effect of a fivefold (relative to the norm) decrease in the ability of the receptors to bind LDL. The calculations demonstrate that the mutation fails to deviate the LDL content in blood plasma and cholesterol concentration in the cell considerably from the norm (**a,** m_2; **b,** m_2), despite a drastic decrease in the LDL transport into the cell from intercellular space through the cell membrane. A compensatory effect of the negative feedback controlling transcription levels of the genes encoding enzymes of the cholesterol biosynthesis and LDL receptor gene in the cell underlies this phenomenon.

A fivefold decrease in the total LDL flow from blood plasma to the cell results in a fivefold increase in the transcription intensities of the corresponding genes according to a negative feedback mechanism. Consequently, both the production of endogenous cholesterol is elevated and the concentration of LDL receptors on the cell surface increases approximately fivefold (**c,** m_2), thereby normalizing the total LDL transport into the cell.

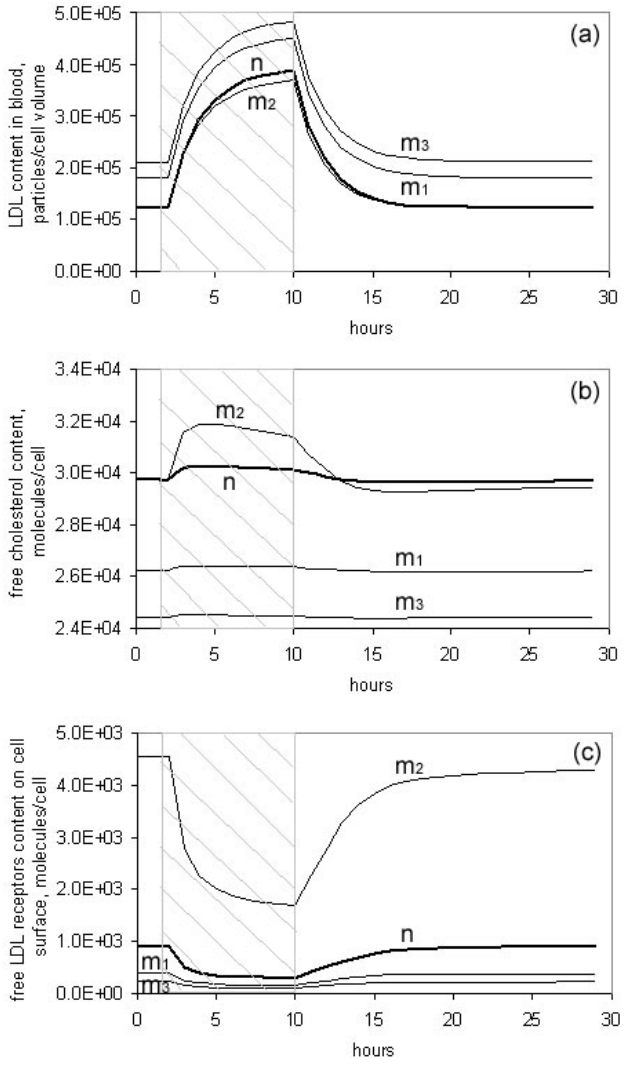


Fig. 2. The responses of the cholesterol biosynthesis gene network in norm (n) and in the presence of various mutations (m_i) to a twofold increase in the LDL inflow into blood plasma continuing over 8 h (hatched region): (**a**) changes in LDL concentration in blood plasma; (**b**) changes in free cholesterol concentration in the cell; (**c**) changes in the number of free LDL receptors on the cell surface; (m_1) the mutation decreasing twofold the LDL receptor synthesis rate; (m_2) the mutation decreasing the binding ability of LDL receptors fivefold; and (m_3) the decrease in the capability of LDL receptors to free LDL in endosomes (a tenfold increase in the receptor degradation rate). The numerical calculations made using the model.

Cleavage of LDL receptor from its ligand in the acid medium of endosomes and its return to the cell surface complete the receptor conversion cycle in the cell. The mutation variant exemplified by curves m_3 (Fig. 2) brings about formation of a truncated LDL receptor protein. This truncated receptor loses the ability to release LDL in endosomes, resulting in receptor degradation. The degradation rates of the LDL receptors impaired by mutations of this class may grow 5–10-fold, decreasing considerably the number of receptors on the cell surface (Fourie *et al.,* 1992). The model allowed us to analyze the response to a tenfold increase in the receptor degradation in the cell relative to the normal rate. Fig. 2 demonstrates that the response to this mutation is qualitatively similar to that caused by the first mutation variants, but is more pronounced. The stationary LDL concentration in blood increases approximately twofold (**a,** m_2), accompanied by a drop in the cell sensitivity to changes in the external LDL concentration.

The stationary number of free receptors on the cell surface reduces approximately 4.5-fold (**c,** m_2); the concentration of free cholesterol, by approximately 25% (**b,** m_2).

Conclusion

The computer analysis of the effects of different mutations in LDL receptor gene on the cholesterol biosynthesis on the cell has demonstrated that the stationary concentration of cholesterol in the cell is sufficiently insensitive to the mutations in question, whereas certain mutations change considerably the LDL level in blood plasma and, correspondingly, the cholesterol level. Certain mutations, such as m_1 and m_3 (Fig. 2) are capable of increasing significantly the LDL concentration on blood plasma, extending their circulation in blood of mutant individuals to 4–6 days (versus 2.5 days in intact individuals). Additionally, these LDLs become more susceptible to chemical modifications (peroxidation, glycosylation, etc.), as it is not native, but these LDL variants that acquire atherogenic properties (Klimov, Nikul'cheva, 1999). Such changes play the key role in development of hypocholesterolemia and atherosclerosis of humans and different animals. Mathematical simulation and computer analysis of the system regulating cholesterol biosynthesis allows new

approaches to prediction of these disease courses to be developed and optimal therapeutic strategies and methods for their corrections to be planned.

Acknowledgements

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

References

1. Brown M.S., Goldstein J.L. (1979). Receptor-mediated endocytosis: insights from the lipoprotein receptor system (Review). Proc. Natl Acad. Sci. USA. 76(7):3330-3337.
2. Brown M.S., Goldstein J.L. (1988). The LDL receptor concept: clinical and therapeutic implications. Atheroscler. Rev. 18:85-93.
3. Fourie A.M., Van der Westhuyzen D.R., Coetzee G.A. (1992). LDL receptor mutations in South African FH patients. Atherosclerosis IX. O. Stein S.Eisenberg, Y.Stein, Eds. Tel Aviv, Israel: R. a. L. Creative Communications Ltd., 153-156.
4. Goldstein J.L., Brown M.S. (1977). The low-density lipoprotein pathway and its relation to atherosclerosis (Review). Annu. Rev. Biochem. 46:897-930.
5. Goldstein J.L., Dana S.E., Faust J.R., Beaudet A.L., Brown M.S. (1975). Role of lysosomal acid lipase in the metabolism of plasma low density lipoprotein. Observations in cultured fibroblasts from a patient with cholesteryl ester storage disease. J. Biol. Chem. 250(21):8487-8495.
6. Hobbs H.H., Russell D.W., Brown M.S., Goldstein J.L. (1990). The LDL receptor locus in familial hypercholesterolemia: mutational analysis of a membrane protein. Annu. Rev. Genet. 24:133-170.
7. Klimov A.N., Nikul'cheva N.G. (1999). Lipid and Lipoprotein Metabolism and Its Disturbances. St.-Petersburg: Piter Kom.
8. Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaya O.A. (2001). A generalized chemical kinetic method for modelling gene networks. Mol. Biol. (Mosk.). 35(6):1072-1079.
9. Likhoshvai V.A., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. (2002). A technology for verifying models of gene network functional dynamics using experimental data. This volume.
10. Murray R.K., Granner D.K., Mayes P.A., Rodwell V.W. (1988). Harpers Biochemistry. Appleton & Lange, Norwalk, Connecticut/San Mateo, California, 1.
11. Russel D.W., Esser V., Hobbs H.H. (1989). Molecular basis of familiar hypercholesterolemia. Ateriosclerosis. Suppl., 9:8-13.
12. Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. (2000). Mathematical model of cholesterol biosynthesis regulation in the cell. Proc. Second on Bioinformatics of Genome Regulation and Structure, Novosibirsk. 1:199-202.
13. Soutar A.K. (1992) Familial hypercholesterolaemia and LDL receptor mutations. J. Intern. Med. 231(6):633-641.
14. Spirin A.S. (1986). Molecular Biology: Ribosome Structure and Protein Biosynthesis. Textbook for Higher Biological Education. M.: Vysshaya Shkola.

CONSTRUCTION OF MATHEMATICAL MODEL OF THE GENE NETWORK ON MACROPHAGE ACTIVATION UNDER THE ACTION OF IFN-γ AND LPS

**** Nedosekina E.A., Ananko E.A., Likhoshvai V.A.***

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

Key words: *mathematical simulation, model, gene networks, macrophage activation*

Resume

Motivation: Activation of a macrophage cell is an important constituent of the immune response. Regulation of this process is supported by the gene network. A necessary stage for theoretical analysis of functioning of this gene network is to simulate an adequate mathematical model. This model will enable to account within the frames of a unique concept the data on structural and functional organization of the gene network, as well as the mechanisms of its particular stages, known values of statistical parameters and dynamic characteristics.
Results: According to the gene network of macrophage activation accumulated in the GeneNet system, we have developed a mathematical model. The current version of the model contains descriptions of 306 elementary processes, with involvement into these processes of 37 genes, about 100 proteins, mRNAs, low molecular weight substances, protein components, etc. By means of the model developed, we have analyzed the action of hypothetic mutations on the gene network functioning.
Availability: The mathematical model is available via the Internet address:
http://wwwmgs.bionet.nsc.ru/systems/MGL/GeneNet/

Introduction

Macrophages are the cells of immune system that are necessary for execution of numerous functions in an organism: resistance to infections, healing of wounds, regulation of functioning of some other types of cells, etc. When infection penetrates into the organism, macrophages become activated, thus, enhancing the synthesis of many substances: enzymes, membrane proteins, low molecular weight substances, etc. The synthesis of these substances is regulated by the gene network of macrophage activation.
Although the macrophage activation process is studied rather intensively, still a lot of details are unknown. In particular, the mediators of many processes are undiscovered; some effects in the gene network regulation are not explained; the quantitative characteristics of molecular processes (e.g., reaction rate constants, substance concentrations in a cell) are obscuredly studied. Among the approaches facilitating solution of some problems mentioned above is the mathematical modelling.
Our goal was to develop a mathematical model of macrophage activation under the action of interferon-gamma (IFN-γ) and lipopolysaccarides (LPS). To this aim, we have used the gene network of this process that was described by applying the GeneNet technology (Kolpakov et al., 1998), which is available via the Internet (http://wwwmgs.bionet.nsc.ru/systems/MGL/GeneNet/) (Nedosekina et al., 2002).

Methods and Algorithms

For simulating the mathematical model, we have used a generalized chemical kinetic simulation method (GCKSM) (Likhoshvai et al., 2001). The GCKSM is based on the block-wise approach. Following this approach, the biological system under study is consequently subdivided into more primitive sub-systems. As a result, the final number of structural elements (mRNA, proteins, low molecular substances, etc.) are unified into a single network by definite elementary processes (reactions of transformation of a substance, regulatory impact, etc.). Each elementary process is described in terms of undependable structural unit of a model. The description is made in terms of standard blocks:

1) Reversible bimolecular reaction: $A + B \underset{k_2}{\overset{k_1}{\leftrightarrow}} C$

$$\frac{dC}{dt} = -k_2 C + k_1 AB = -\frac{dA}{dt} = -\frac{dB}{dt}$$

2) Non-reversible monomolecular reaction: $A \xrightarrow{k} B_1 + B_2 + \ldots + B_n$

$$\frac{dA}{dt} = -\frac{dB_1}{dt} = -\frac{dB_2}{dt} = ... = -\frac{dB_n}{dt} = -kA \ , n \geq 0.$$

3) Constitutive synthesis: $\xrightarrow{k} B_1 + B_2 ... B_n$

$$\frac{dB_i}{dt} = k, i = 1,..., n, n \geq 1.$$

4) Enzymatic reaction: $E + S \underset{k_2}{\overset{k_1}{\leftrightarrow}} ES \overset{k_3}{\leftrightarrow} E + P$

$$\frac{dP}{dt} = \frac{V_{\max}[S]}{K_M + [S]}, \text{ where } V_{\max} = k_3[E]_{total}, K_M = \frac{k_2 + k_3}{k_1}$$

As a result of simulation, we arrive at the formalized description of the processes of gene network functioning, or the structure-functional model.

Then this structure-functional model is automatically transformed into the mathematical form by applying the software program «Model designer» developed previously (Likhoshvai et al., 2000). If the external stimuli are absent, this mathematical form corresponds to the system of the ordinary differential equations, with dynamical variables expressing concentrations of genes, mRNA, proteins, low molecular weight substances and their complexes. In general case, the mathematical model is compiled of differential equations and discrete expressions.

The key stage under construction of a model is the stage of verification of its parameters. To this aim, we have ordered initial approximation values of the model parameters on the basis of analysis of literature data. Then we have performed the search for the optimal values of these parameters. The task was solved numerically on the basis of the novel technology (Likhoshvai et al., 2002). This technology is oriented on developing the databases, which accumulate the data on dynamics and scenarios of the model's functioning under different external impacts.

The search for the optimal model parameters is realized by original software program, which applies the method for simulation of the evolution of a population of individuals. Each individual is considered as a career of the simulated system with individual set of parameters' values, which are generated occasionally. At each stage in evolution, the best fitting individuals are being selected. The fitness is understood in terms of similarity of the results obtained by the model to the experimental data published.

Results and Discussion

The current version of the mathematical model contains descriptions of 306 elementary processes, which include 37 genes, about 100 proteins, mRNAs, low molecular weight substances, protein complexes and intermediate substances.

Currently, we have developed scenarios on 12 experiments, which were devoted to studying of 9 gene networks' components: NO, MHCII IA^b (mRNA and protein), MHCII IE^b mRNA, CIITA mRNA, IP-10 mRNA, TLR4 mRNA, IL-12p40 protein, and TNF-α protein. They enabled us to use at the stage of the model's adaptation 46 experimentally estimated concentration values of the substances mentioned above and measured at definite periods of time.

An example of comparison of the results of calculations, obtained after adaptation of the model, to the experimental data are shown in Figures 1 and 2.

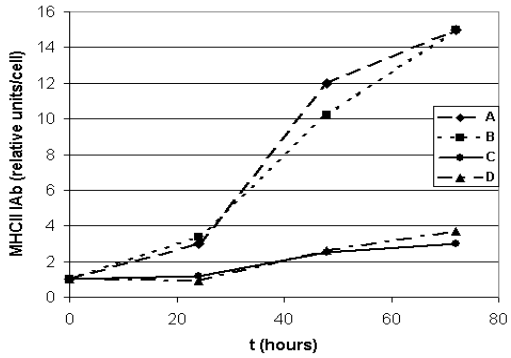


Fig. 1. Dependence of MHCII IAb mRNA and protein concentrations upon the period of IFN-γ influence. A, mRNA, experiment (Herrero C. et al., 2001); B, mRNA, model; C, protein, experiment; D, protein, model.

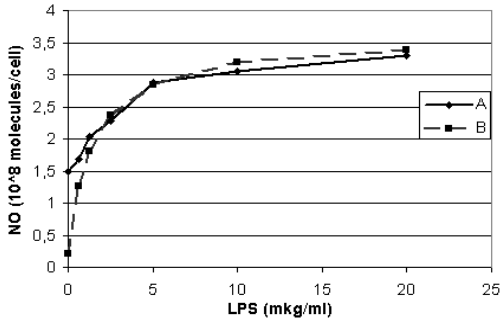


Fig. 2. Dependence of elaborated NO amount upon the LPS concentration: A, experimental data (Chen F. et al, 1995); B, model simulation

We have applied the mathematical model constructed to analysis of dynamics of the gene network functioning under supposition about the presence in it of a series of mutations. In Fig. 3, the calculation of the impact of hypothetical mutations in two genes, *iNOS* (encoding the enzyme named inducible nitric oxide synthase) and *GCH1* (encoding the enzyme GTP cyclohydrolase I, which catalyses the first stage of tetrahydrobiopterin co-factor synthesis), on NO production is illustrated. We suppose that hypothetical mutations decrease 10-fold the turn-over constant, Ko, of enzymes iNOS and GCH1. As follows from Fig. 3, in 24 hours after LPS induction, the NO production is decreased 10-fold and 4-

fold, respectively (compare pairwise the curves A and B, A and D in Fig. 3). We have performed a numerical analysis of the most natural supposition claiming that compensatory renewal of NO production rate under the action of the mutant enzyme iNOS could be achieved by the ordinary increase in the level of the basal transcription of the *iNOS* gene. However, even 4-fold increase in the rate of mRNA transcription produced no significant compensatory effect on the level of NO production (Fig. 3, curve C). This example demonstrates that due to non-linear effects in the complex gene networks, the search for effective strategies compensating the action of mutation is a non-trivial task.
In Fig. 4, an example of the positive solution of the task of searching for compensatory impact is shown. We have considered a mutation that enhances 10-folds the expression efficacy of the CD14 receptor, which is, in its turn, being the LPS receptor.

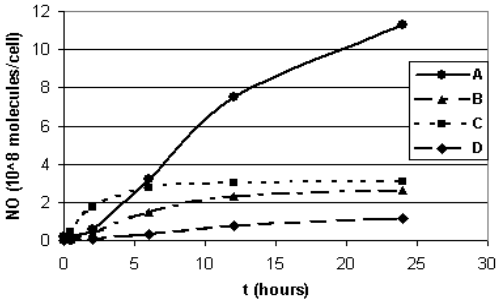


Fig. 3. Alteration of NO concentration under the action of LPS: **A,** in norm; **B,** under mutation, decreasing 10- fold K_o of GCH1; **C,** under mutation decreasing 10-fold Ko of the enzyme iNOS, in case the basal transcription of the *iNOS* gene is increased by 10^4 times; **D,** under the action of mutation decreasing the Ko value of the enzyme iNOS by 10 times.

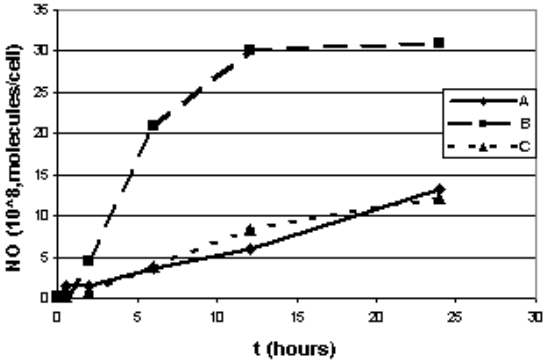


Fig. 4. Dynamics of NO synthesis under LPS induction. A, experimental data (Chen et al., 1995); B, influence of the over-expression of CD14, model calculations; C, compensatory effect of inputting a substance, which bounds the excess of CD14 , model simulation.

Such mutation leads to significant NO overproduction (Fig. 4, compare pairwise the curves A and B). Due to this phenomenon, different pathologies could appear, in particular, septic shock. Among the impacts compensating the action of such a mutation, could be the binding of the excess of receptor by some substances (Fig. 4).
By numerical modelling, we have selected the optimal concentration of this hypothetical substance and the constant of its binding/disintegration to the CD14 receptor.

Implementation

The model presented in this paper describes the macrophage activation under the action of LPS and IFN-γ. We have performed adaptation of the model parameters to the experimental data. The current version of the model enables to calculate dynamic characteristics of the gene network components under different external stimuli. The model could be applied to solving some topical tasks including: a) the search for optimal characteristics of the substances, which exert onto the gene network *a priory* known impact; b) testifying the alternative hypotheses about the missing links and/or regulatory relationships; c) studying the impact of mutations and characteristics of the gene network behavior under different conditions; d) solving the tasks of optimal stimuli for compensating the disruptions of the gene network functioning.

Acknowledgements

The authors are grateful to Alexander Ratushny for help in constructing the model. The work was supported in part by the Russian Foundation for Basic Research (№ 01-07-90376, 00-04-49229). Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Project № 65).

References

1. Chen F., Sun S.C., Kuh D.C., Gaydos L.J., Demers L.M. (1995) Essential Role of NF-kB Activation in Silica-Induced Inflammatory Mediator Production in Macrophages. Biochem. and Biophysical Res. Commun. 214, 985-992.
2. Herrero C., Marques L., Lloberas J., Celada A. (2001) IFN-gamma-dependent transcription of MHC class II IA is impaired in macrophages from aged mice. J. Clin. Invest. 107, 485-493.
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. Likhoshvai V.A., Matushkin Iu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. (2001) A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk). 35, 1072-1079.
5. Likhoshvai V.A., Matushkin Yu.G. Vatolin Yu.N., Bazhan S.I. (2000) A generalized chemical kinetic method for simulating complex biological systems. A computer model of λ phage ontogenesis. Comp. Technologies. 5, 87-99.

6. Likhoshvai V.A., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. (2002) Technology of usage of experimental data for verification of the models of gene network functioning. Proc. III Intern. conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
7. Nedosekina E.A., Ananko E.A. (2002) Gene network of macrophage activation under the action of interferon-gamma and lipopolysaccarides. Proc. III Intern. conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

ANALYSIS OF MUTATIONAL PORTRAITS OF GENE NETWORKS

**** Ratushny A.V., Likhoshvai V.A., Kolchanov N.A.***

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

Key words: *computer analysis, gene network, mutational portrait*

Resume

Motivation: Development of optimal strategies for correcting various pathologies taking into account the genotype-specific distinctions of particular individuals and detection of the targets for pharmacological regulation require the knowledge on behavior of the system in question under various conditions as well as in the case the rates of its constituent processes have changed. Analysis of mutational portraits of gene networks is an approach allowing this knowledge to be obtained.
Results: The sensitivity of free cholesterol stationary content to mutational changes in the rates of molecular processes running within the corresponding gene network was analyzed. It was demonstrated that the mutations hitting regulatory processes changed the free cholesterol stationary content to a largest degree.
Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/

Introduction

Gene networks are groups of concertedly expressed genes that control the vital functions of the body (Kolchanov et al., 2000). Gene networks may comprise dozens, hundreds, and even thousands of components with a most intricate structure–function organization. In the majority of cases, these objects display a nonlinear behavior due to their intrinsic negative and positive feedbacks. Any random mutational change in the rate of a process within a gene network may switch on a cascade of changes in the overall biological system, probably resulting in development of a pathology. Thus, the gene network function requires correction to either reconstitute its normal parameters or bring them to the values closest to the norm. Study of the qualitative and quantitative behavior patterns of the system with changed rates of biochemical processes within the gene network in question is a necessary background for developing the strategies allowing the function of this biological system to be corrected, including therapeutic correction.
In this work, a "mutational portrait" of the gene network regulating cholesterol biosynthesis in the cell and its exchange with blood plasma cholesterol is "drawn".
The mutational portrait of a gene network is regarded as a set of the stationary states and dynamic characteristics of the gene network obtained through varying "mutationally" the rates of all its constituent elementary processes within specified ranges.
The sensitivity of free cholesterol stationary content in the cell to mutational changes in the rates of molecular processes within this gene network was analyzed. It was demonstrated that the mutations hitting the regulatory processes change the free cholesterol stationary content to a largest degree.

Methods and Algorithms

A computer model of function dynamics of the gene network regulating cholesterol biosynthesis in the cell (Ratushny et al., 2000) was used to construct and analyze its mutational portrait. The model, developed within the framework of generalized chemical kinetic approach (Likhoshvai et al., 2001), comprises 82 elementary processes and is described with 39 common differential equations and 97 constants (http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/). Verification of the parameters of this model is detailed in (Ratushny, Likhoshvai, 2002). Additional information is available in (Ratushny et al., 2000) and the papers used to describe this gene network in GeneNet (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/).

Implementation and Results

Effects of individual mutations of varying intensity of all the 97 constants used in the model and characterizing the processes running in the gene network were studied. The "mutated" values of a constant was considered equal to 13 values of the geometric series with a step of 2 where the first term was equated to the initial value of the constant in question divided by 2^7. The stationary state of the gene network was calculated upon each "mutation"; overall, 1165 calculations were made.
Typical curves reflecting the changes in free cholesterol stationary content in the cell depending on mutational changes in the four following parameters of the model are shown in Fig. 1: (1) exchange constant of the enzyme SRP (sterol regulated protease) changed approximately twofold; (2) constant of the reverse reaction of SREBP1 (sterol regulatory element-

binding protein) dimerization changed by 65–70%; (3) Michaelis–Menten constant of the enzyme acetoacetyl-CoA thiolase, by 15–20%; and (4) exchange constant of the enzyme ACAT (acyl-CoA: cholesterol acyltransferace) did not change at all.

Fig. 1. Changes in free cholesterol stationary content in the cell upon mutational changes in parameters of mathematical model of the gene network regulating cholesterol biosynthesis: (1) exchange constant of the enzyme SRP (sterol regulated protease); (2) constant of the reverse reaction of SREBP1 dimerization; (3) Michaelis–Menten constant of the enzyme acetoacetyl-CoA thiolase; and (4) exchange constant of the enzyme ACAT (acyl-CoA: cholesterol acyltransferace).

The results of the analysis performed are summarized in Fig. 2. Mutational changes in the intensities of the processes marked with double exclamation point (!!) exert largest effects of the free cholesterol stationary concentration (from 0 to over 200% relative to the norm). All these processes are involved in the functioning of regulatory circuits of the gene network in question. The mutations hitting these processes impair the corresponding regulatory mechanisms and, consequently, affect the cholesterol stationary concentration in the cell. Activation of the key component of this gene network—a transcription factor SREBP1 (framed in Fig. 2)—is an example of such processes; other examples, the processes controlling the activation rate. However, the fraction of the processes constituting the gene network of cholesterol biosynthesis that exert considerable effects on intracellular cholesterol stationary concentration is not large, amounting to approximately 15%.

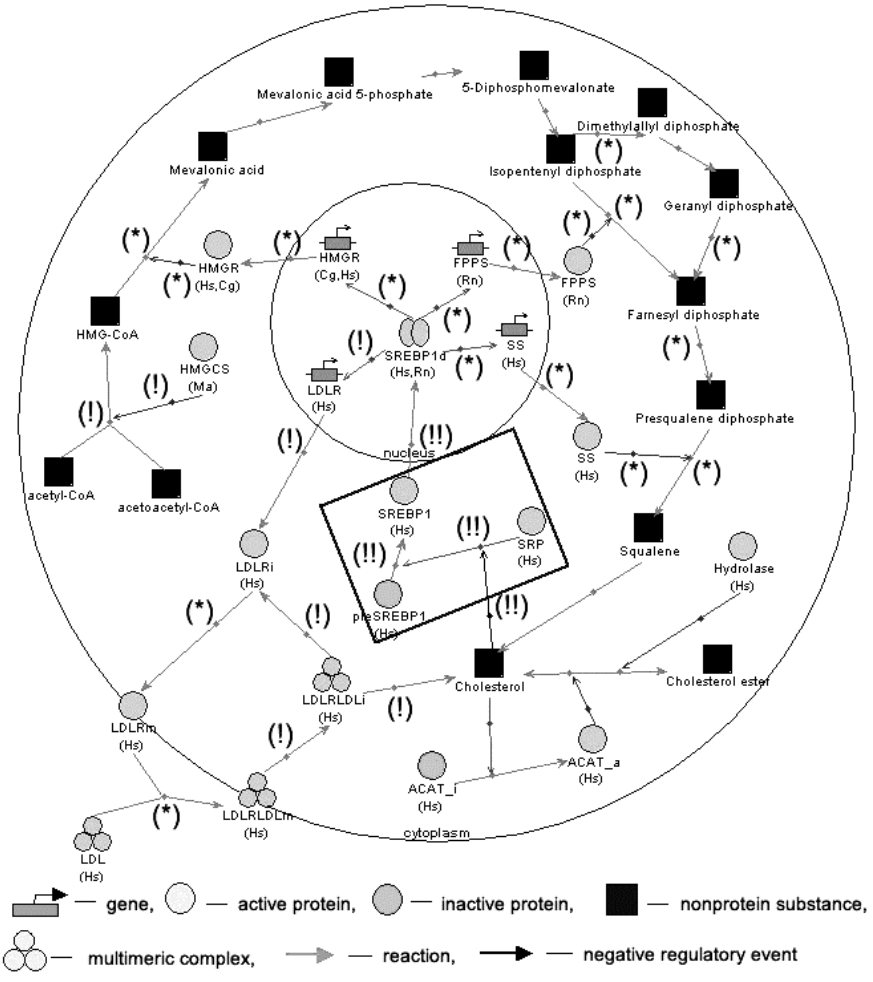


Fig. 2. Gene network of cholesterol biosynthesis in the cell. The elementary processes found to be sensitive to mutations by the analysis performed are designated as follows: (!!) the cholesterol stationary concentration changes by 0 to 200% relative to the norm upon mutational changes in the rates of the corresponding processes (one example is framed) in the range of ± 2 orders of magnitude; (!) the changes are below 35% of the norm; and (*) the changes are below 25%.

The processes marked with single exclamation point (!) are involved in the initial stages of cholesterol synthesis in the cell and its intake. Mutational changes in the rates of these processes influence the cholesterol stationary content varying it less than by 35% with regard to the norm.
The changes in the rates of the processes marked with asterisk (*) are even smaller, changing the cholesterol free concentration less than by 25%. These processes are mainly involved in cholesterol synthesis and, in part, recirculation of LDL receptors.
Mutational changes in the rest processes failed to affect the cholesterol stationary concentration in the cell.
The analysis performed has demonstrated that the free cholesterol stationary concentration determined by the gene network in question is indifferent towards the mutational changes in the majority of parameters of this gene network. Mutational changes in 85% parameters of the mathematical model in the ranges considered fail to influence the cholesterol stationary concentration or vary it not more than by 35% relative to the norm.
What underlies the discovered weak sensitivity of cholesterol concentration to mutational changes in the rates of the majority of processes constituting this gene network? The following circumstances appear essential in this connection. (i) The presence of non-limiting stages in the biochemical pathways of the network in question. (ii) Occurrence of two processes responsible for appearance of additional cholesterol amounts in the cell, namely, (a) cholesterol biosynthesis in the cell *per se* and (b) cholesterol transport from blood plasma into the cell via LDL receptors. Mutation-caused impairment of either process brings about a compensatory effect due to its counterpart. Thereby, the cell has a reliable double regulation of free cholesterol inflow. (iii) Shunting of certain biochemical reactions within the cholesterol biosynthesis pathway may also play similar role in elevation of mutation stability of this gene network. Finally, (iv) of the highest importance is the negative feedback regulation of the intracellular cholesterol concentration.

Conclusion

Tremendous number of molecular genetic, biochemical, and physiological processes, controlled by gene networks, proceed simultaneously in cells, tissues, organs, and bodies. Analysis of the information compiled with the GeneNet database and in the available published sources allowed us to distinguish four major types of gene networks: (i) gene networks controlling irreversible processes, such as cell growth and differentiation, morphogenesis of tissues and organs, growth and development of organisms; (ii) gene networks regulating cyclic processes, such as cell cycle, heart muscle contraction, etc.; (iii) gene networks involved in homeostasis of biochemical and physiological parameters of the body; and (iv) gene networks of body responses to environmental changes, for example, stress response (Kolchanov et al., 2000).
The results obtained are typical of the gene networks of type (iii), where negative feedbacks play the key role. As for the rest types of gene networks, mutations may influence their function according to qualitatively different patterns.
Study of gene network mutational portraits is of special importance while searching for optimal targets for pharmacological regulation. Values of the constants of certain reactions occurring in the body may be changed with the available pharmacological tools even now. Undoubtedly, the future potential in correcting the functions of gene networks though a directed change in their particular constituents will be extended, making the theoretical analysis of gene network function in general and gene network mutational portraits in particular an essential element of novel biotechnological approaches.

Acknowledgements

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

References

1. Likhoshvai V.A., Matushkin Yu.G., Ratushnyi A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaya O.A. (2001). A generalized chemical kinetic method for modelling gene networks. Mol. Biol. (Mosk.). 35(6):1072-1079.
2. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko E.L. (2000). Gene networks. Mol. Biol. (Mosk.). 34(4):533-544.
3. Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. (2000). Mathematical model of cholesterol biosynthesis regulation in the cell. Proc. II Intern. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk. 1:199-202.
4. Ratushny A.V., Likhoshvai V.A. (2002), Computer analysis of the effect of mutations in LDL receptor gene on the system regulating the intracellular cholesterol biosynthesis. This issue.

EVOLUTION OF DIPLOID GENE NETWORK OF CHOLESTEROL BIOSYNTHESIS REGULATION IN A CELL

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

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

Key words: *evolution, computer model, diploid gene network, cholesterol biosynthesis*

Resume

Motivation: The main problem of the evolution theory is the transition between micro- and macroevolution. More and more data support the evidence that the novel forms appear as a result of mutations in regulatory genes. Alteration in regulation modifies the whole gene network determining a trait. Thus, we may conclude that evolution of gene networks determines formation of species.
Results: We have studied the model task of simulating evolution of diploid gene network on biosynthesis regulation in a cell. As was shown, adaptation may touch on a single locus or many loci simultaneously in dependence upon the stringency of alterations of external environment. In the course of this process, adaptation concerns mainly the loci with mutations producing slightly favorable/deleterious or quazi-neutral effects. The loci responsible for functioning of non-limiting stages of gene networks, as well as the loci with mutational alterations exerting strong damaging effects are not crucial for selection.
Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/gn_model/

Introduction

Evolution of gene networks becomes one of the pivotal problems in the theory of evolution. As was repeatedly shown, even weak alteration in the gene network regulation may exert profound effects on the organism-carrier. Even more strong effects could be produced by appearing novel regulatory contours in a gene network (Kolchanov, Matushkin, 1997). From this point of view, rather perspective is the direct studying of the models of evolution of gene networks with the portrait modelling of the occurring mutations, especially almost neutral ones, which influence could be hardly estimated "by eye".
In the work presented, we simulate the functioning in evolution of the gene network of cholesterol regulation in a cell under transition from the normal environment to the environment with the low influx of low density lipoproteins (LDL) in an organism. The simulation was made by modification of the computer model developed by us previously (Ratushny et al., 2000). In this modification, we have accounted for diploidy of genes. We have simulated the evolution of isogeneous cell strains under the action of mutations and selection directed for adaptation to novel environmental conditions. Based on the results obtained, we have analyzed the functioning of hybrid gene networks obtained by crossing the gene networks, which are evolutionally adapted to environments with different intensity of LDL supply in the blood plasma.

Methods and Algorithms

In the current work, we have used the variant, accounting for diploidy, of the mathematical model that enables to study the functioning of a gene network on cholesterol regulation in a cell and its metabolism with the cholesterol of the blood plasma. The description of the model is given elsewhere by A.V.Ratushny and co-authors (Ratushny et al., 2000).
The procedure of simulation of evolutionary process:
In the course of simulating the evolution, we have supposed that each individual is a carrier of the modeled gene network with the individual set of parameters. This individual may give rise to the fixed number of offspring who may differ from the initial ancestral individual by one or several mutations, which are randomly generated. At each stage of evolution, the selection samples the particular number of the best fitting individuals. The fitness is understood in terms of the relatedness of results obtained by the model simulations to the experimental data published. The conditional functional of adaptability, W, is calculated as follows:

$$W = \frac{F_N}{F}, \text{ where } F = \sum_i \left(\frac{x_i^t}{x_i^e} + \frac{x_i^e}{x_i^t} - 2 \right).$$

Here x_i^t is the i-th value, obtained by numerical calculations of the mathematical model of the gene network (for example, the concentration of a definite gene network's component under particular conditions); x_i^e is the i-th experimental value; F_N is the value of F in the norm.
The conditional functional of adaptability was built on the basis of the experimental data described in publications (Brown, Goldstein, 1979; Goldstein et al., 1977; Goldstein et al., 1975).

Implementation and Results

At the initial moment of the evolutionary time, it is supposed that all the individuals in population are the carriers of the gene networks adapted to the normal conditions (environment N), proposing that the rate of LDL influx into the blood plasma equals to V_N. In this case, W_N=1. Then the individuals were replaced into the environment L, in which the rate of LDL influx into blood was set as the half of the norm: $V_L=V_N/2$. In the environment L, their conditional adaptability decreased to W_L=0.004. This means that the organisms, which were initially adapted to the environment N, are characterized by the low adaptability in the environment L.

In the course of evolutionary adaptation of the gene network in homozygous condition to the environment L, the value of the functional was W_L=0.75. The meaningful result is that although evolutionary alterations of the gene network were allowable for all the parameters, the final variant was selected, which solved the problem of evolutionary adaptiveness due to fixation of mutations sharply increasing transcription of a gene coding the low density lipoprotein receptor. That is, only a single component of a gene network was altered. Thus, the problem of adaptation in evolution was solved due to elaboration of the more effective mechanism of transportation of lipids into the cell from the low-grade lipid environment. Notably, only those mutations were fixed in evolution that altered only a single locus of a gene network, which was responsible for the synthesis of LDL receptors (Fig. 1(1)). As a result of evolution, the rate of their synthesis decreased by 2.3-fold relatively the norm, whereas the adaptability of the gene network to the environment with the low influx of LDL into the blood has risen from 0.004 to 0.75. The constants of the rates of the other processes were almost invariable.

Next, we have simulated an evolutionary adaptation of the gene network, adapted to the environment L, to novel conditions, which are characterized by even more low influx of LDL into blood plasma (four-fold decrease in comparison to the norm and two-fold in comparison to the environment L). As a result, the high adaptability to novel conditions of the external environment was achieved due to alteration of 11 parameters of the system, which refer to description of 6 processes shown in Fig. 1(2). Hence, in this case, the only recourse of variation in the LDL receptors synthesis rate is insufficient for gaining high adaptability.

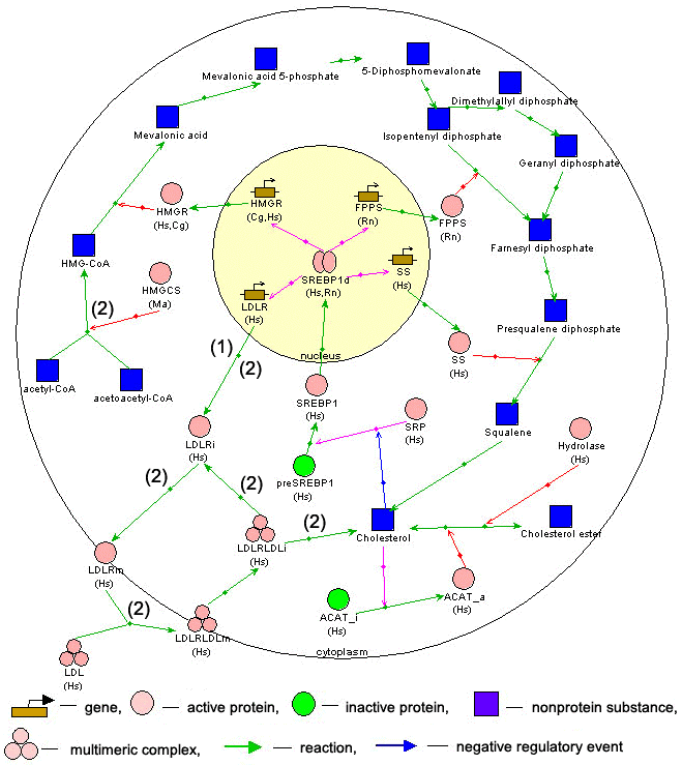


Fig. 1. Gene network on cholesterol biosynthesis regulation in a cell. Elementary processes, which were changed in the course of evolutionary adaptation of the network to the environment with the decreased influx of LDL are denoted as follows: 1, the rate of LDL influx into blood was set two-fold less than in norm; 2, the four-fold decrease, respectively.

Notably, thee processes that were involved in evolutionary adaptation, both in the first and the second cases, refer, due to our data (Ratushny, 2002), to the factors slightly limiting the gene network on cholesterol biosynthesis in a cell. In other words, although their alterations by mutations change the cholesterol level in a cell, these changes are not of dramatic character. Thus, under conditions of the constant environment, mutations in these loci exert slightly deleterious (or quazi-neutral) action. However, exactly these mutations provide a possibility for a diploid gene network on cholesterol synthesis in a cell to adapt in evolution to a novel environmental conditions (when alterations caused by mutations in these loci became of clear adaptive value).

As was noted, the loci responsible for functioning of non-limiting stages of the gene network (in this case, selection is ineffective at all), as well as the loci with mutations exerting disrupting effects (here selection is ineffective due to considerable decrease in adaptability), were not involved in the process of evolution.

Thus, our result gives evidence about important potential in evolution of slightly deleterious and quazi-neutral mutations, which favor to adaptation.

Modelling of Crosses

In the work presented we have analyzed the results of crossing of diploid gene networks that are adapted to the environment N with the normal rate of LDL influx into the blood, as well as to the environment L with decreased support by LDL. Three variants are possible: *AA,* homozygous condition of the LDL receptor system, for which W_N=1 in the environment N; aa, homozygous condition of the LDL receptor system, for which W_L=0.75 in the environment L; Aa, heterozygous condition of the LDL receptor gene.
The Table illustrates the values of conditional functional of adaptability W for different conditions of biological system of LDL receptor gene in various environments.

Table. Conditional functional value under different conditions of the gene network.

	W_L	W_N
AA	0.004	1
Aa	0.03	0.92
aa	0.75	0.70

The interesting and unexpected result is that the individuals with the gene network with the *aa* condition of the LDL receptor gene, which are adapted to the low-grade lipid environment (W_L = 0.75), possess by rather high adaptability to the environment with the normal level of lipids concentration (W_N = 0.70), analogously to individuals with the initial gene network *AA* (W_N =1).
This means that the individuals *aa* adapted to the environment with the low-grade concentration of a substrate are characterized by more broad norm of adaptation rather than initial individuals *AA*, with W_L = 0.004. Clearly, in case migration is possible between two habitats, N and L, the organisms with the gene networks *aa* could gradually assimilate the habitat N, thus, providing the individuals *AA* with the serious competition. The organisms with the hybrid gene network *Aa* also possess by good adaptability to the environment N (W_N = 0.92). Summing up, we may conclude that the environment N, as a whole, supports the *A* allele, whereas the environment L favors the *a* allele. The exchange between environments leads to equilibrium between the A and *a* alleles, with predominance of the *A* over the *a* allele in the environment N, and, on the contrary, possibly more contrast pattern of allele distribution in the environment L.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (№ 01-07-90376, 02-07-90359), Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Project № 65).

References

1. Brown M.S., Goldstein J.L. (1979) Receptor-mediated endocytosis: insights from the lipoprotein receptor system. Proc. Natl Acad. Sci. USA. 76(7), 3330-3337. Review.
2. Goldstein J.L., Brown M.S. (1977) The low-density lipoprotein pathway and its relation to atherosclerosis. Ann. Rev. Biochem. 46, 897-930. Review.
3. Goldstein J.L., Dana S.E., Faust J.R., Beaudet A.L., Brown M.S. (1975) Role of lysosomal acid lipase in the metabolism of plasma low density lipoprotein. Observations in cultured fibroblasts from a patient with cholesteryl ester storage disease. J. Biol. Chem. 250(21), 8487-8495.
4. Kolchanov N.A., Matushkin Yu.G. (1997). A Biological Self-reproducing System: Principles of Organization and Evolution. Russ. J. of Genet. 33(8), 889-897.
5. Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. (2001) A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk). 35(6), 1072-1079. (Russian).
6. Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. (2000) Mathematical model of cholesterol biosynthesis regulation in the cell. Proc. of the second Intern. conf. on bioinformatics or genome regulation and structure. Novosibirsk. 1, 199-202.
7. Ratushny A.V., Likhoshvai V.A. Computer analysis of the action of a mutation in LDL receptor gene on the system of cholesterol regulation in a cell. This issue.
8. Ratushny A.V., Likhoshvai V.A., Kolchanov N.A. Analysis of mutational portraits of gene networks. This issue.
9. Likhoshvai V.A., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. (2002) Technology of experimental data application to verification of the models of dynamics of gene network functioning. This issue.

DIAGNOSTICS OF MUTATIONS BASED ON ANALYSIS OF GENE NETWORKS

Borisova I.A. [1]***, * Zagoruiko N.G.*** [2]***, Likhoshvai V.A.*** [3]***, Ratushny A.V.*** [3]***, Kolchanov N.A.*** [3]

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

Key words: *mutation, pattern recognition, pairwise comparison, space of patterns, erythroid differentiation*

Resume

Motivation: Diagnostics of mutations in biological systems is one of the most important problems in present-day genetics. Rapid development of novel technologies, including laboratories-on-a-chip, produces simultaneous qualitative dynamic characteristics of the most biological molecules in a cell. Combination of these data with the data obtained by the Data Mining approach makes the task of mutation recognition rather feasible.
Results: The method for pattern recognition based on consecutive comparison of unlabeled object with the standards of pairs of patterns is described. The comparison of competing hypotheses is made in the subspace of features chosen for recognition of each pair of patterns separately. An example of recognition the types of single and double mutations in the gene network of erythroid differentiation is resulted. The high self-descriptiveness of the dynamic data about change of concentration of the biologically relevant molecules in time is routined.
Availability: http://www.math.nsc.ru/AP/oteks

Introduction

The task of mutation recognition (diagnostics) in the gene networks is one of the most actual problems in modern theoretical genetics. This problem becomes feasible against the background of the novel experimental technologies, which have appeared recently and attain large-scale distribution. These technologies automatically reveal dynamic characteristics of cell functioning of hundreds and thousands of genes and their products. Among such technologies are biochips that are useful for analysis of dynamics of alterations of gene transcription (DNA Chips, Gene Chips), protein synthesis, protein interaction (Protein Microarrays, Proteome Chips), and even of low-molecular cell processes (Khandurina, Guttman, 2002). For analysis of such tremendous bulk of experimental evidence on dynamics of a system's behavior, a question arises about approaches aimed at treatment of these data, their theoretical analysis, and practical usage. As the most important task of analysis of this sort of data may be viewed the task of searching for mutations in biological systems and recognition of their type. Realization of this task will enable to develop the methods of diagnostics of diseases caused by disruption in a gene network functioning, as well as pharmaceuticals with the narrow spectrum of action on preordered molecular-genetical and biochemical cell processes, etc.

Methods and Algorithms

Simulation of mutations. In this work, we have studied the gene network on erythroid cell regulation under the action of erythropoietin. Description of this biological system and the respective mathematical model are given previously (Ratushny et al., 2000). An additional information could be extracted from publications, which lay in the basis of description of this gene network in the database GeneNet (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/).
The model is constructed within the frames of generalized chemical kinetic approach (Likhoshvai et al., 2001). Using the model, it is possible to observe dynamics of this gene network behavior under various external impacts. By exploiting the model, the data were obtained on alterations of concentrations of various substances acting in biochemical reactions. The observation was made, during 100 transit hours, of alterations in dynamics of 34 substances, including mRNA and enzymes of the heme synthesis, the heme itself, α- and β-globins and the relevant mRNAs, hemoglobin, key regulator of erythrocyte differentiation, transcription factor GATA-1, some cell receptors, etc.
Mutations disrupting the functioning of a particular stage in a gene network were simulated by varying this or that parameters of the model. All in all, we have simulated 9 different types of mutations. In Figures 1-3, one can see dynamics of concentrations of (1) heme, (2) receptors bound on the cell surface with the transferrin, and (3) GATA-1 mRNA in erythroid cell (a, in norm; b, under the action of mutations). The kinetics observed, of maturation of the «mutant erythrocytes», together with calculations for the system in norm, were used as the training sample. Each realization executes the role of the typical representative (precedent) of its image.

For each mutation, the kinetic curves were noised by random-number generator in the limits 20-30% from initial number. In such a manner, 10 variations were generated for each mutation. The resulted sample, containing 90 realizations, was used for recognition of the single mutations.

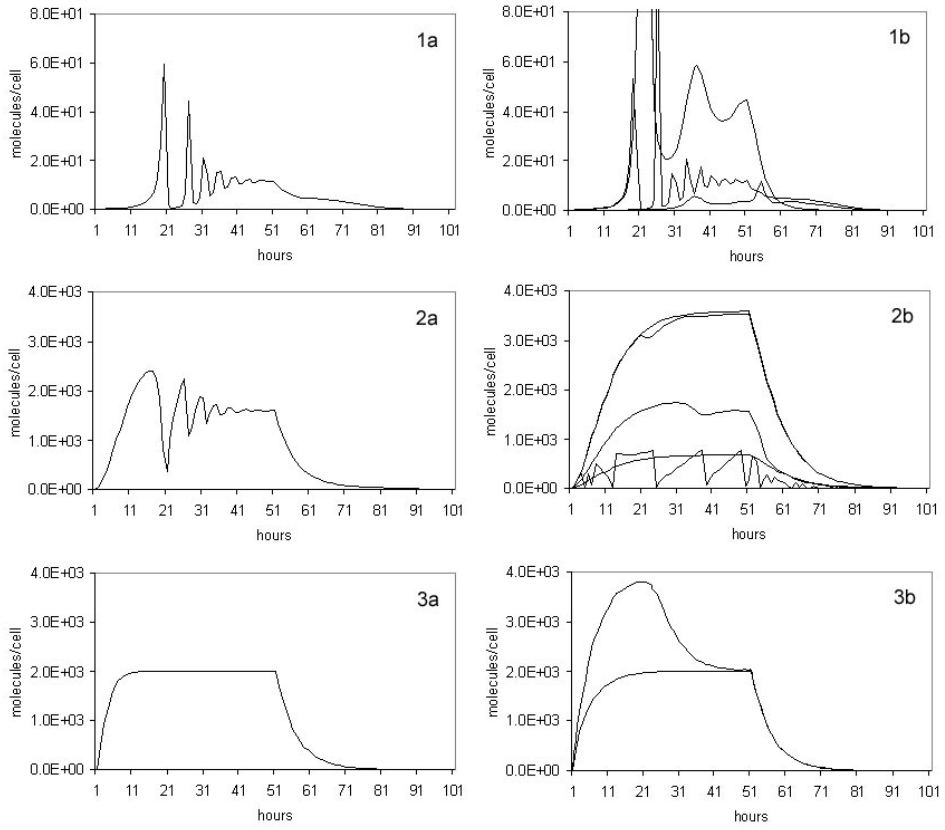


Fig. Dynamics of changes in concentrations of (1) heme, (2) receptors bound at the cell surface with transferrin, and (3) GATA-1 mRNA in erythroid cell (a, in norm; b, under different mutations). Zero point of time, the cell-precursor; 80 hours, mature erythrocyte.

Recognition method. Recognition of a large number of patterns is usually performed in some feature space common for them. Obviously, for reliable recognition of pairs of patterns (A, B), (A, C), and (B, C), it would be expedient to use the most competent features, individually picked up for each pair. In such feature space, the patterns of each pair will most strongly differ from the others. For example, by recognising the oral words "sixteen", "fifteen", and "fifty", for the first pair of words, it is necessary to take into account the features related to beginnings of these words, while for the second pair, these features should deal with the ends of words. If K patterns are distinguished by this approach, it is required to construct the standards for all the pair combinations and then during recognition to organise economical procedure for pairwise comparison of competing patterns. Let's begin with a choice of the most competent subspace of features for each pair of patterns.

Selection of the competence subspace. Each realisation of any K pattern is described by an initial set of N features. If the feature Xi accepts the same value both for a pattern A and pattern B, then we need to consider its competence for this pair as equalling to 0. If the distance between average values (m) of this feature for the patterns given is great, this feature may be informative. Additionally, we need also to take into account dispersion (d) values of this feature over all training objects of each pattern. Such approach to estimation of potential competence J of features is well produced by the Fisher's criterion: $J = |m(A)-m(B)| / (d(A)+d(B))$.

Let's accept competence (Ji) as a weight factor of an i-th feature at definition of weighed distance between objects in a feature space. As a result, for each pair patterns, the N-dimension space is considered through its individual variant of weight factors. Each this variant of the weighed space we shall name as "competent subspace" X(A,B) for a given pair patterns A and B. Generally, by comparing an object Y with the patterns A and B, the distances R(a,Y) and R(b,Y) in the subspace X(A,B) will be determined through the weighed Euclidean distance (Zagoruiko, 1999).

T*he method of pairwise comparison.* Let us recognise whether an object Y belongs to one of three patterns: A, B, or C. By comparing the object Y to the standards of patterns A and B, we determine to which of these two patterns Y is more similar in a subspace, which is optimal exactly in this case. Analogously, comparison between the patterns of the other pairs, (A, C) and (B, C), is made. If in two pairwise comparisons (A, B) and (A, C), the pattern A becomes the winner, then the winner among the patterns B and C should not win the pattern A. Thus, it is unnecessary to compare B and C.

Based on the evidence given above, the recognition procedure is the following. At the training stage, we form a single standard for each of K patterns, as well as (K-1) lines with N weights of competence factors for each pair of patterns. The number of lines of competence equals to K*(K-1)/2. The N-dimension feature vector of the object Y is compared to the standard vectors of any pair of patterns by using a line of factors for this pair. The pattern is determined in such a way that the weighed distance between its standard and the object Y is minimal. This pattern occupies the first position, while the second position is taken by any pattern, which was not participating in competition yet. The winner of this pair again occupies the first position, whereas the next possible competitor is put at the second position. This procedure is repeated (K-1) times.

If the winner of the last couple became a leader in the very first comparison, then the process is finished. If not, it should pass a comparison with the patterns to which it was not compared yet. As a result, decision whether the object Y belongs to one of K patterns is made approximately after K steps. Thus the destiny of each pattern is determined in the most favourable conditions. This important and natural requirement is not valid in traditional methods of recognition in feature space, common for all patterns.

In general, the recognition algorithm suggested contains (i) the stage of choosing competent features for all pairs of patterns and (ii) the stage of deleting the weakest competitors by comparing the control object Y with the pairs of standards. The efficiency of this algorithm was checked up for several applied tasks. One of them is related to studying gene networks.

Implementation and Results

Recognition of single mutations. For all 45 pairs from 10 patterns, we have estimated the competency of 34 substances, including those that could be measured easily. As experiments have shown, the information applicable for mutation recognition (kinetic curves for 34 substances) has a great redundancy. It turned out that for exact recognition of all the objects, information about only 3 gene network components is needed: heme, receptors, bound at the surface with transferrin, and GATA-1 mRNA (see Figures 1-3). For each substance in 100 moments of time, for each pair of compared hypotheses, is suffice to use only 20 moments of time.

According to the most informative characteristics chosen, all the control mutations were recognized without mistakes. Due to this fact, it is possible to develop a very economical and rapid approaches aimed at diagnostics of the type of single mutations.

Recognition of double mutations. By means of the same mathematical model of the gene network, we have obtained the data on a gene network's behavior for double mutations, which were formed under the action of all possible pairs of 9 single mutations considered above. These 36 types of double mutations were recognized in accordance with the same 60 features (concentrations of three substances during 20 moments of time), which were used for recognition of single mutations.

The program gave an output with the most probable single mutations constituting the double mutations. In the Table, one may see the comparison between contents of 36 double mutations and the results of their recognition.

Table. Comparison of double mutations obtained by mathematical model of a gene network to results of their recognition.

Type of double mutation	1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 3 3 3 4 4 4 4 4 5 5 5 5 6 6 6 7 7 8 2 3 4 5 6 7 8 9 3 4 5 6 7 8 9 4 5 6 7 8 9 5 6 7 8 9 6 7 8 9 7 8 9 8 9 9
Result of recognition	1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 3 3 **4 4** 3 3 4 4 4 4 4 **8** 5 5 5 6 6 6 7 7 8 2 **4** 4 5 6 7 8 9 3 4 5 6 7 8 9 4 5 6 7 8 9 5 6 7 **7** 9 6 **3** 8 9 7 **9 9** 1 **4** 9

The recognition was considered as correct in case real single mutations occupied two first positions in the list. Correct recognition was obtained in 89% of cases. The results of the experiment verify the high reliability of recognition method based on deleting the weakest competitors under pairwise comparison in competent subspace of features.

Conclusion

By applying the pattern recognition, we have detected a significant redundancy in characteristics of dynamic characteristics in a gene network. The principal possibility of single mutation recognition, as well as recognition of double mutations in dependence upon penetration characteristic for single mutations. This approach clears the way to development of diagnostic methods recognizing diseases caused by disruption of a gene network functioning.

Acknowledgements

The work was supported in part by the Russian Foundation of Basic Research (№ 01-07-90376, 02-07-90359, 02-01-00082), Russian Ministry of Industry, Sciences and Technologies (№ 43.073.1.1.1501), Siberian Branch of the Russian Academy of Sciences (Integration Project № 65).

References

1. Khandurina J., Guttman A. (2002) Bioanalysis in microfluidic devices. J. Chromatogr. A. 943(2), 159-183.

2. Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. (2001) A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk). 35(6), 1072-1079. (Russian).
3. Ratushny A.V., Podkolodnaya O.A., Ananko E.A., Likhoshvai V.A. (2000) Mathematical model of erythroid cell differentiation regulation. Proc. of the 2nd Intern. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk. 1, 203-206.
4. Zagoruiko N.G. (1999) Applied Methods for Data and Knowledge Analysis. Ed. of Institute of Mathematics SD RAS, Novosibirsk. 268.

ON THE THEORY OF PREDICTION OF GLOBAL MODES IN THE FUNCTION OF GENE NETWORKS

Likhoshvai V.A., Matushkin Yu.G.

Institute of Cytology and Genetics, SB RAS, 630090, Novosibirsk, Russia, e-mail: likho@bionet.nsc.ru
*Corresponding author

Key words: *gene network, mathematical model, computer model, regulation, negative feedback, positive feedback, critical points, limit cycle, stability*

Resume

Motivation: The actual gene networks are intricately arranged, and their structures are yet vague. Understanding of their function requires not only further studies of naturally occurring gene networks, but also distinguishing of their standard elements and research into their interactions in different combinations within the concept of theoretical constructs. Study of the properties of theoretical constructs will help in understanding the function patterns of natural gene networks.
Results: Theoretical objects—hypothetical gene networks constructed from two types of elementary units, genetic elements and regulatory links—are introduced in this work. A criterion allowing all the stable points of four important classes of hypothetical gene networks with negative feedback regulation to be computed basing on the structure of their oriented graphs is formulated.

Introduction

The live systems display most sophisticated internal structure. The ability to self-reproduction and efficient existence under changing environmental conditions are the unique features of these systems. Gene networks, providing performance of the vital functions, play an exclusively important role in the functioning of live systems. Usually, gene networks comprise dozens and hundreds of elements of different nature and complexity: genes and their regulatory regions; preRNAs, mRNAs, and proteins, encoded by these genes; low-molecular-weight compounds; various complexes between proteins and their targets, etc. The elements of a gene network are integrated into a single entity through complex nonlinear processes (Kolchanov et al., 2000).
The genes encoding regulatory proteins together with negative and positive feedbacks regulating the gene activities play the most important, if not the determining, role in the function of gene networks. It is their presence that renders the gene networks capable of self-regulating and reacting adequately to changes in the environmental conditions (Kolchanov, 1997). Although the understanding of the importance of negative feedbacks in the regulation of ontogenesis goes back to Jacob and Monod (1961) and the effects of negative and positive feedbacks of the function dynamics of gene networks were theoretically analyzed earlier (Thomas et al., 1995), their roles in establishing the global properties of gene networks require further studies. We are considering the global properties of gene networks as their stable states at any permissible values of their internal parameters and external conditions. As the problem stated in its general form is yet hardly subjectable to constructive analysis, we followed the path of decreasing the diversity of the study subjects. We are introducing a new theoretical object—hypothetical gene networks (HGNs). We will construct HGNs of two major type elements—genetic elements and regulatory links. This approach was applied in (Likhoshvai et al., 2001) and opens the possibility of structural analysis of actual gene networks. Further development of this work, described here, allows solving of the practically important problem of constructing gene networks with prespecified dynamical properties and stable states to be approached.
A criterion allowing the global properties of four HGN classes, such as the presence or absence of stable points, to be described in a certain parameter ranges is formulated. The criterion is constructive, based on the analysis of the properties of gene network structure graph, and requires no calculations of HGN function dynamics.

Implementation and Results

The actual gene networks comprise a finite number of substances and processes. Genes, mRNAs, proteins, their various forms and intermediate complexes, low-molecular-weight compounds, etc., represent substances. Processes integrate the substances into a single entity—the gene network. Changes in concentrations of substances characterize the function of gene network in the time domain. Biochemical processes and the processes of active and passive mass and energy transfer form the basis of the gene network function. Therefore, differential equations are applicable to mathematical description of functional dynamics of gene networks. If the assumption on an instant mixing (uniformity of distribution of the substances) is admissible, the simulation can be made in terms of the following ordinary differential equations:

$$dx_i/d\,t = F_i(X_{i'},K) - x_iG_i(X,K),\ i = 1,\ldots,n, \qquad (1)$$

where n is the number of dynamic variables of gene network model; $X = (x_1,\ldots,x_n)$, the vector of dynamic variables; $K = (k_1,\ldots,k_m)$, the vector of parameters; $X_i' = (x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)$, the vector of the dynamic variables lacking x_i; and $F_i(X_i;K)$ and $G_i(X,K)$, rational functions taking nonnegative values at all the nonnegative values of variables and parameters.

Thus, it is possible to consider the study of the operation patterns of gene networks as a study of the systems of type (1). However, the systems of type (1) form a very wide subclass, dense within the class of ordinary differential equations. Consequently, the general statements attributable to these systems are very scanty. In particular, there is a general theory (Gorban et al., 1986) answering the question on global limit properties of type (1) systems. Development of such theory becomes now essentially more topical due to intensive research into the structure of natural gene networks and demands of pharmacogenetics.

In this work, we are considering hypothetical gene networks (HGNs) that are constructed of two types of elements—genetic elements (GEs) and regulatory links (RLs). We consider GE as an idealized object that includes gene, mRNA, and the protein encoded and implicitly, all the elements and processes providing their syntheses and formation of their active forms (Fig. 1a); RL, as an idealized process decreasing/increasing GE activity via the corresponding regulator (Fig. 1b). The GE activity manifests itself in the protein (regulator, R) synthesis rate. Multimers, which are generally formed of Rs encoded by several GEs, are the active form of regulators (active regulators; ARs). Let us construct a HGN from a finite number of GEs through interconnecting them with a certain number of regulatory links. Let us describe the properties of the HGN wherein n Rs are synthesized using systems of ordinary differential equations of a special type (2), where concentrations of monomeric proteins p_i are the dynamic variables; the negative term determines the rate of R degradation, and Z_i determines the overall mechanism regulating activity of the ith R synthesis:

$$\frac{dp_i}{dt} = Z_i - \beta_i p_i \quad i = \overline{1,n}\,. \tag{2}$$

The overall mechanism regulating the activity of R synthesis is composed of elementary events (EEs) according to certain rules. Totally, we distinguish four EEs. A certain rule for composing the formal equation is ascribed to each EE (Table 1). Combining the EEs, listed in Table 1, hypothetical gene networks with a very complex regulation patterns can be constructed.

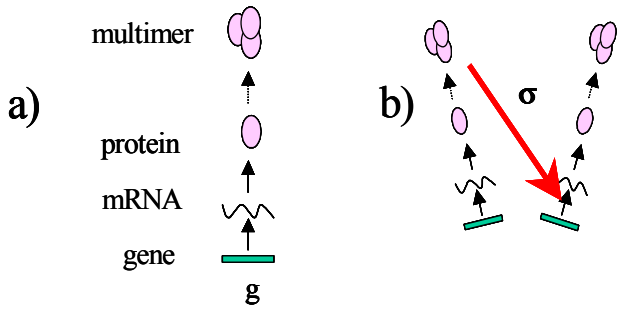


Fig. 1. Elementary units of HGN.

Table 1. Types of elementary events.

Type of elementary events	Rule for composing formal equation	Comments
Rs form AR in solution.	$\prod p_{i_u}^{\gamma_u} = p_{i_1}^{\gamma_1}\ldots p_{i_k}^{\gamma_k}$	$i_1,\ldots,i_k$ are numbers of Rs forming AR; $\gamma_1,\ldots,\gamma_k$ are degrees of multimerization.
Several ARs affect the same regulatory site with a conditional number v.	$S_{i,v} = \dfrac{\alpha_{i,0} + \sum \alpha_{i,j} \prod p_{i_u}^{\gamma_u}}{\delta_{i,0} + \sum \delta_{i,j} \prod p_{i_u}^{\gamma_u}}$	The equation for $S_{i,v}$ is composed of only those ARs that affect the site in question; if $\delta_{i,j}\alpha_{i,j} \neq 0$, AR activates the site; if $\delta_{i,j} \neq 0$, $\alpha_{i,j} = 0$, inhibits.
GE with a conditional number j has several nonoverlapping regulatory sites.	$A_{i,j} = \prod S_{i,v}$	
Several GEs encode one protein.	$Z_i = \sum A_{i,j}$	i is the number of protein.

Then, let us impose additional limitations on the diversity of elements used for constructing individual HGNs (Table 2). This allows us to distinguish four classes of hypothetical gene networks. Let us correspond oriented graphs (OGs) to these HGNs in a one-to-one manner and formulate the hypothesis on the number of stable points.

Table 2. Limitations imposed on construction of individual HGNs.

Limitation	Description
1	Only negative feedbacks are considered.
2	All the GEs encode different proteins.
3	Each GE has only one regulatory site.
4	All ARs are homomultimers formed by proteins of one type.
5	Only one AR corresponds to each GE; all ARs are heteromultimers formed by proteins of different types.
6	All the ARs affect one site of GE.
7	Each AR acts through individual GE site distinct from those for other ARs.

Class 1: HGNs are constructed using limitations 1, 2, 3, 4, and 6;
Class 2: HGNs are constructed using limitations 1, 2, 4, and 7;
Class 3: HGNs are constructed using limitations 1, 3, and 5; and
Class 4:. HGNs are constructed using limitations 1, 2, 3, and 5.

For **classes 1** and **2,** the HGN structure–function organization and OG structure correspond in a one-to-one manner, if oriented graph nodes correspond to HGN GEs, while the OG edges reflect inhibition of the GE located at the sink node by the protein multimer encoded by the GE located at the source node. For **class 3,** the GEs are oriented graph nodes, while all the edges coming to a particular node reflect the formation of the AR inhibiting the GE located at this sink node. For **class 4,** a one-to-one correspondence means that each OG node corresponds to particular protein, while each edge corresponds to GE encoding the protein located at the sink node and is repressed by the protein located at the source node.

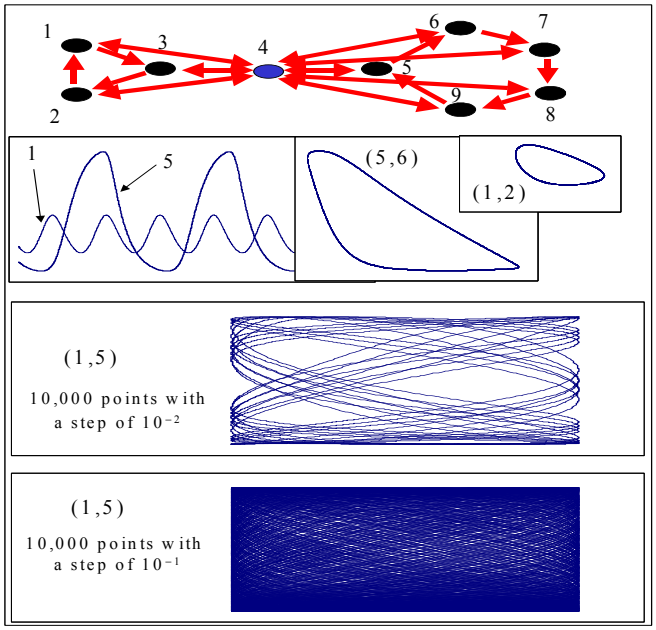


Fig. 2. An example of HGN constructed of nine genetic elements and calculation of its dynamic behavior at $\gamma = 5$, $\alpha = 3$ in the region of continuous oscillation.

Fig. 2 exemplifies numerical calculations of a **class 1** HGN. This HGN is constructed of nine GEs. It has a region of continuous oscillations; there are good grounds suggesting that they are quasiperiodic.
In conclusion, let us formulate a hypothesis that allows all the stable points to be calculated for any HGN of the four classes introduced within a certain **range of their parameters. Let us further assume that $\alpha_{i,0} = \alpha$ $\delta_{i,j} = 1$ and $\gamma_{i,j} = \gamma$, $\beta_i = 1$ (see Table 1 and equation 2).**

Criterion

Let an oriented graph G be given. Let S(G) denote a HGN of a certain class constructed using the oriented graph G. Then, γ_0 and α_0 exist for S(G), such that the number of stable points for this HGN equals the number of 1-bases of the oriented graph G at any $\gamma > \gamma_0$ and $\alpha > \alpha_0$.
If each node of G belongs to at least one 1-base (see the definition in Harary, 1973), S(G) has no other stable states.

Conclusion

Thus, we have introduced hypothetical gene networks and studied their properties. A criterion relating the global limit properties of four classes of hypothetical gene networks to properties of their structural oriented graphs is formulated; namely, the criterion allows the HGN stable points to be calculated at certain, rather high values of the parameters γ and α. Finally, when applied to actual gene networks, occurrence of a certain gene network structure (oriented graph) is a

necessary condition for existence of a specified number of their stable points. The realization requires a minimal complexity (nonlinearity) of the processes regulating activities of the gene network genetic elements. The required complexity may be achieved through an increase in the degree of multimerization of repressor proteins and/or the occurrence of sufficiently high number of intermediate stages leading to formation of repressor proteins.

Acknowledgements

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

References

1. Gorban A.N. Bykov V.I., Yablonskii G.S. (1986) Essays on Chemical Relaxation, Kiperman S.L. (Ed.). Novosibirsk: Nauka.
2. Harary F. (1973). Graph Theory. Gavrilova G.P. (Ed.). M.: Mir.
3. Jacob F., Monod J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. J. Mol. Biol. 3:318-356.
4. Kolchanov N.A. (1997). Transcription regulation of eukaryotic genes: databases and computer analysis. Mol. Biol. (Mosk.). 31:581–583.
5. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000). Gene networks. Mol. Biol. (Mosk.). 34:449-460.
6. Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2001). On connection of a graph of a gene network with qualitative modes of function. Mol. Biol. (Mosk.). 35(6):1080-1087.
7. Thomas R., Thieffry D., Kaufman M. (1995). Dynamical behavior of biological regulatory networks. I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state. Bull. Math. Biol. 57:247-276.

A STUDY OF THE FUNCTION MODES OF SYMMETRIC GENETIC NETWORKS

Likhoshvai V.A. [1]****, Matushkin Yu.G.*** [1]***, Fadeev S.I.*** [2]

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

Key words*: genetic networks, hypothetical genetic networks, mathematical model, computer model, regulation, negative feedback, positive feedback, particular points, limiting cycles, stability*

Resume

Motivation: Analysis of the influence of the structure of genetic networks on their function characteristics is an important problem of bioinformatics. The study of the characteristics of theoretical constructions is necessary to an uderstanding of the function patterns of natural genetic networks.
Results: This paper analyzes the symmetric hypothetical genetic networks (SHGN) regulated by negative feedback. The (n,k)-criterion is formulated. It can be used to compute without calculations all stable function modes for SHGN and to determine whether these are stable points or stable limiting cycles. This can be performed only by analyzing the divisibility of number n by number k (n- the number of genetic elements in the network), (k-1) – the number of regulators of an individual genetic element).

Introduction

Practical needs of the analysis of the dynamics of genetic network behavior call for numerical investigation of the corresponding mathematical models (see, e.g. Gardner et al., 2000, Elovich, Leiber, 2000). In this case, it is necessary to answer some questions that can be considered standard. Among these are the problems on the qualitative modes of genetic network behavior for the different values of parameters, the determination of stable points, the nondecaying function modes of either cyclic or more complex nature, the establishment of interrelation between the structurally functional organization of a genetic network and all probable function modes, the study of the problems of stability (parametric, structural, evolutionary, etc.). The answers to these questions are not only of theoretical but also of practical interest. For example, these are essential for the construction of genetic networks with preassigned characteristics, the development of expression vectors with novel characteristics, the determination of the optimum strategies of governing genetic networks, the development of problems related through the needs of pharmaco-genetics to the construction of biocomputers.
To comprehensively analyze the function pattern of genetic networks, we have introduced a theoretical object, i.e., hypothetical genetic networks, HGN (Likhoshvai, Matushkin, 2002). The HGNs consist of the elements of two types: genetic elements and regulatory relationships. The HGNs allow one to study the role of negative and positive feedbacks in the formation of the global characteristics of genetic networks neglecting the concrete individual peculiarities of their structure and the conditions of natural genetic networks function. The latter is of importance because in many cases, the structure of natural genetic networks is still only partially understood. The present report considers the characteristics of symmetric hypothetical genetic networks. A numerical study shows that their limiting characteristics are fully determined by two numbers: n – the number of genetic elements in the network, k-1 – the number of regulators of an individual genetic element. We formulate the (n,k)-criterion for describing the global portrait of symmetric HGNs function from the values of n and k. In this case, there is no need to calculate the function dynamics of HGNs.

Implementation and Results

The hypothetical genetic networks (HGN) as theoretical constructions whose characteristics should be studied for creating the function theory of genetic networks are presented in (Likhoshvai et al., 2001) and futher developed in (Likhoshvai, Matushkin, 2002). The HGNs consist of the elements of two types: genetic elements (GE) and regulatory relationships (RR). Their characteristics are described by the sets of standard differential equations of a particular form where the concentrations of p_i proteins in a monomeric form are the dynamic variables. This study is focused on the symmetric HGNs that are constructed as follows. The two natural numbers are fixed $2 \leq k \geq n$. Then we take n genetic elements and arrange them in sequence by cycle. Thereafter it is assumed that any gene is the inhibitor of the following k-1 genes clockwise. For example, if k=2, the first gene inhibits the second one, the second gene inhibits the third one, etc., the n-th gene inhibits the activity of the first gene, at k=3 the first gene inhibits the following two genes, etc. by cycle (Fig. 2). For the genetic network of n genes, there is the n-1 symmetric genetic network without autoinhibition. As n and k unambiguously determine symmetric networks, we called them the S(n,k) networks and the corresponding structural digraphs and the models introduced in (Likhoshvai, Matushkin, 2002) were denoted by G(n,k) and M(n,k), respectively.

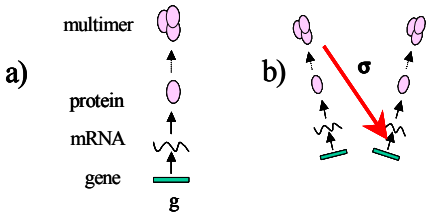


Fig. 1. HGN elementary units.

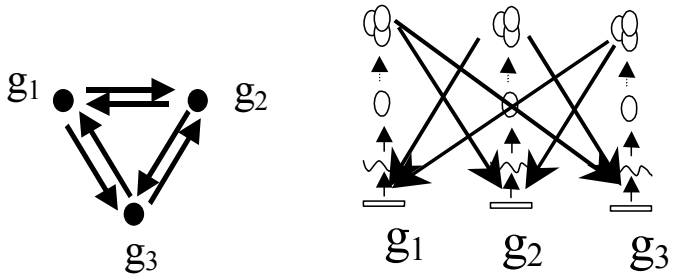


Fig. 2. The scheme of SHGN with three genetic elements and the corresponding digraph.

The M(n,k) models for the symmetric HGNs of the first class (for definition see Likhoshvai, Matushkin, 2002), under constraints $\alpha_{i,0}=\alpha$, $\delta_{i,j}=1$, $\gamma_{i,j}=\gamma$, $\beta_i=1$, are of the form

$$dp_i/dt = -p_i + \alpha/(1+p^{\gamma}_{\mathrm{mod}_n(i-1)}+\ldots+p^{\gamma}_{\mathrm{mod}_n(i-k+1)}),\ i=\overline{1,n}\ , \quad (1)$$

where $\mathrm{mod}_n(i)=\begin{cases} i, если\ 1\le i\le n \\ i-n,\ если\ n<i\le n+k-1 \end{cases}$

For the other classes, the systems are readily derived from the rules given in (Likhoshvai, 2002).
The second-class HGNs obey the following systems of equations

$$dp_i/dt = -p_i + \alpha/(1+p^{\gamma}_{\mathrm{mod}_n(i-1)}\cdot\ldots\cdot p^{\gamma}_{\mathrm{mod}_n(i-k+1)}),\ i=\overline{1,n}. \quad (2)$$

Those of the third class are described by the systems

$$dp_i/dt = -p_i + \alpha/\left[(1+p^{\gamma}_{\mathrm{mod}_n(i-1)})\cdot\ldots\cdot(1+p^{\gamma}_{\mathrm{mod}_n(i-k+1)})\right],\ i=\overline{1,n}. \quad (3)$$

For the fourth class, the equations are of the form

$$dp_i/dt = -p_i + \alpha\left[1/(1+p^{\gamma}_{\mathrm{mod}_n(i-1)})+\ldots+1/(1+p^{\gamma}_{\mathrm{mod}_n(i-k+1)})\right],\ i=\overline{1,n}. \quad (4)$$

The numerical calculations show that depending on the structure, the symmetric HGNs can either have the stationary points or function as continuous oscillators. In this case, some HGN constructions can have more than one stable function modes. We have numerically studied the symmetric HGNs with two to nine genetic elements. The results were the same for all classes except for the second one (see the Table). For the second class, the differences were recorded only for HGN variants with continuous oscillations (lines 11 and 12 in the Table). In this case, only one limiting cycle was revealed for each variant.

Table. The list of the numerically studied M(n,k)-models and the number of stable particular points (P) and limiting cycles (C) found from them at $\alpha>5$, $\gamma>k$.

№	n	k	P	C
1	2,4,6,8	2	2	0
2	3,6,9	3	3	0
3	4,8	4	4	0
4	n=k	5,6,7,8,9	k	0
5	3,5,7,9	2	0	1
6	4,5,7,8	3	0	1
7	5,7,9,	4	0	1
8	6,7,8,9	5	0	1
9	7	6	0	1
10	8	6		2
11	6	4	0	2 (1*)
12	9	6	0	3 (1*)

* The number of limiting cycles found for second-class HGNs.

Analysis of the calculated results for the models given in the Table and those omitted in this paper allows the following generalizing hypothesis.

The (n,k)-criterion. For any fixed n and k there are γo and αo such that with any $\gamma \geq \gamma o$ and $\alpha \geq \alpha o$ one and only one of the following statements is fulfilled: 1) if n is totally divided by k, the M(n,k) models of any class have k stationary points and no other stable modes; 2) if n cannot be totally divided by k, the M(n,k) models of the first, second and third classes have d (d being the greatest common divisor of n and k) stable periodic modes and the second-class model has a single stable cyclic mode of behavior and no other function modes are observed in this region).

The (**n**,**k**)-criterion is observed to establish a simple rule of elucidating the limiting characteristics of the symmetric hypothetical genetic networks without calculations. We can just compare only two numbers, n and k. It is noteworthy that the characteristics predicted by the (n,k)-criterion hold only if γ and α exceed some limits depending, generally speaking, on n, k, and the HGN class. When these conditions are not fulfilled, the HGN can have another limiting portrait. Thus, with great γ and α, the qualitative behavior is completely determined by the structure of the symmetric hypothetical genetic network and with small γ and α, of importance are the concrete values of these parameters. Note also that in the section concerning the counting of stable points, the (n,k)-criterion results from a more general criterion given in (Likhoshvai, Matushkin, 2002). However, in the section devoted to the determination of the structures of symmetric HGNs with limiting cycles and to the counting of the cyclic modes of behavior, the (n,k)-criterion provides additional information that cannot be extracted from the previous criterion. It is assumed then that both of the criteria are the special cases of the more general criterion for estimating the global characteristics of HGN.

Conclusions

The result given in this paper describes the global characteristics of the symmetric hypothetical genetic networks. We assume that developing further the theory of hypothetical genetic networks will make it possible to approach the solution of the practically important problem, i.e., the construction of genetic networks with preassigned dynamic characteristics and limiting function modes (the number of stationary and/or oscillating variants of dynamic behavior). The fact that a qualitative behavior of the genetic network depends on the graph structure and can cardinally change with either the appearance of at least novel or the disappearance of the previous regulatory relationship, offers new means of both explaining the patterns of genetic networks evolution (in particular, their evolutionary complication) and interpreting the influence of mutations on the genetic networks function and the processes under their control.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (№ 02-04-488802, 01-07-90376, 02-07-90359), Russian Ministry of Industry, Sciences and Technologies (№ 43,073.1.1.1501), Siberian Branch of Russian Academy of Sciences (integration Projects № 65).

References

1. Gardner T.S., Cantor C.R., Collins J.J. (2000) Construction of a genetic toggle switch in *Escherichia coli*. Nature. 403, 339-342.
2. Elowitz M.B, Leibler S. (2000) A synthetic oscillatory network of transcriptional regulators. Nature. 403, 335-338.
3. Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2001) Relationship between a Gene Network Graph and ualitative Modes of Its Functioning. Mol. Biol (Mosk). 35, 1080-1087.
4. Likhoshvai V.A, Matushkin Yu.G. (2002) On the theory of prediction of the global function modes of genetic networks. Proc. I Intern. conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

DEVELOPMENT OF THE PROGRAM SOFTWARE FOR MATHEMATIC MODELLING OF THE GENE NETWORK DYNAMICS

[1] Fadeev S.I., [1] Berezin A.Yu., [1] Gainova I.A., [1] Kogai V.V., [2] Ratushny A.V., [2] Likhoshvai V.A.*

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

Key words: *mathematical modelling, numerical analysis, continuation by the parameter*

Resume

Motivation: Analysis of mathematical models of biological systems is an essential stage for their further development. To this aim, it is necessary to develop special methods and the program software.
Results: In this work, GeneNetSTEP software complex containing the STEP package and BPR-Q package, oriented to numerical studying of the autonomous systems of differential equations are presented.

Introduction

The study of the structure-functional organization and dynamics of gene networks is proceeding vigorously, thus making development of tools for computer-assisted visualization of dynamic characteristics of the gene networks a foreground task. Among important problems is how to analyze the models developed. In its turn, to fulfil this task, it is necessary to develop appropriate software. In this work, we consider the methods of numerical modelling of mathematical models, in which the functioning of the gene networks is described by autonomous systems of equations of the form:

$$dX_i/dt = F_i(X,K) - X_i*G_i(X,K), \quad i = 1,2,\ldots N, \qquad (1)$$

where X is a vector of concentrations of substances of a gene network, K is a vector of inner parameters of the gene network processes, F_i (X,K) and G_i(X,K) are the rational functions, which describe the laws of alterations of a substance concentration (Likhoshvai, Matushkin, 2002; Likhoshvai et al., 2001a). The study of behavior of the solution (1) in dependence upon the parameters is an important constituent of the general problem of developing mathematical models adequate to experimental data. To the same class of problems refer the studying of mathematical models of hypothetical gene networks, which are the particular cases of (1) (Likhoshvai et al., 2001b). In this case, the second members of the system of equations have a very simple special view, which preserves qualitative description of the laws regulating activities of expression of genetic elements in hypothetical gene networks. Due to this fact, it is possible to predict, what sorts of limiting solutions, that is, solutions of (1) in case $t \to \infty$, are typical for the model considered, as well as to estimate analytically a number of the characteristic properties of these solutions (Likhoshvai et al., 2002).
A software complex represented in this work is oriented to the numerical study of the autonomous systems of the form (1). They comprise the algorithms enabling to study numerically the abstract autonomous system of N equations of the form:

$$dy/dt = f(y,\alpha), \qquad (2)$$

where α is one of the model's parameters. Also, they include the integrated systems (2) with set initial conditions (Cauchy problem), searching for stationary solutions in dependence upon the parameter α, or for solutions of the system of nonlinear equations

$$f(y,\alpha) = 0, \qquad (3)$$

and analysis of stability of the stationary solutions obtained. Previously, these algorithms were arranged as a software package STEP, which is widely applicable for the studying of kinetic equations of various catalytic processes (Fadeev et al., 1998). For studying behavior of the limiting cycles in dependence upon the parameters, the algorithms for numerical studying of nonlinear boundary-value problems for the systems of ordinary differential equations are included into the GeneNetSTEP software complex, namely, software package BPR-Q (Kogai et al., 2001). Integrally, the algorithms enable to make a numerical experiment aimed at studying the properties of nonlinear problem solutions of (2), (3), in rather general problem setting on the basis of the method of continuation by a parameter.

Results

In this work, we present the GeneNetSTEP software complex, adapted to analysis of the models of gene network functioning. To this aim, we have developed a program-converter for the automated input of mathematical models of the gene networks developed within the frames of the generalized chemical kinetic simulation method (Likhoshvai et al., 2000; Likhoshvai et al., 2001b).

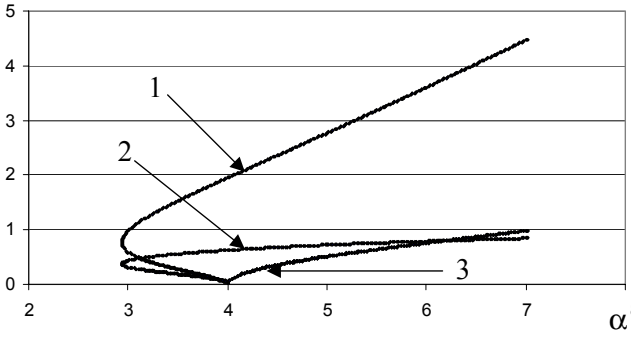


Fig. An amplitude of oscillation of variables in the model M(6,4), for γ=4. The first stable cycle: variables x1, x3, x5, the amplitude 1; variables x2, x4, x6, the amplitude 2; the second stable cycle: variables x1, x3, x5, the amplitude 2; variables x2, x4, x6, the amplitude 1; the third stable cycle, all the variables have the amplitude 3. By abscissa, the parameter α from the system of equations (4); by ordinate, the amplitude of oscillation of variables of the model M(6,4).

For increasing efficacy of searching for Cauchy problem solutions within the frames of the software package STEP, we have additionally accounted for a peculiarity of the gene network models, that is, the strong sparseness of the matrix of derivatives. Under the large dimensionality of the system (1), accounting for this peculiarity in realization of the computer algorithms related to solving the systems of linear equations is an important condition of their efficacy. In this connection, on the basis of ideas represented in the monograph by S.Pissanetsky (1988), the modification of the Gauss's elimination method was developed and realized as a software program aimed at solution of this sort of systems. The programs are included into the software package STEP as an alternate variant for the classic Gauss's elimination method for the systems with the sparse matrices.

The software package STEP consists of four sections. In the first section, the construction of mathematical model and the bank of models are presented. In the second section, the STEP package contains the semi-implicit Rosenbrock method of the 2-nd order and multistep Gear algorithm for integration of stiff ordinary differential equations. In the third section, the solution of the system of nonlinear equations is given in dependence upon the parameter α according to the method of continuation by the parameter. To determine stability of the stationary solutions, the numerical κ-criterion by Godunov-Bulgakov is used in the package. This criterion is based on effective method on determining the norm of the matrix H of solutions of the matrix Lyapunov equation: $HA + A^*H=I$, where $A=f_y(y,\alpha)$. The fourth section is devoted to interpretation of results, i.e., the results given in a form of tables, plots, to searching for functions in dependence upon a solution, and to storing of information in a file form.

For numerical test analysis of the software package STEP, we have used a mathematical model of dynamics of a gene network on cholesterol biosynthesis regulation in a cell and its exchange with cholesterol from blood plasma (Ratushny et al., 2000). The basic stages of this gene network are quantitatively accumulated in the database GeneNet (http://wwwmgs.bionet.nsc.ru/systems/mgl/genenet/). The mathematical model is based on 82 elementary processes; it contains 39 dynamic variables, 97 reaction constants, and 21 parameters, which express concentrations of non-variable components of the gene network (genes, precursors of other elements, etc.). It was numerically estimated that the model has a single equilibrium state in a broad spectrum of values.

It is of importance to reveal and analyze the continuous trajectory of the system of equations of the form (1). In the complex of the software programs BPR-Q, continuous waves are studied by the method of the continuation by the parameter. At every continuation step, for solving the nonlinear boundary problem, we use quazi-linearization on the basis of multiple shooting method by using differential sweep method (Kogai, Fadeev, 2001). It is important to note that the methods of the BPR-Q enable to search for unstable cycles, which are known to be undetectable by general methods.

As an example, let us consider the model M(n,k) of a hypothetical gene network of the class 1 (Likhoshvai et al., 2002) with the parameters $\alpha > 0, \beta > 0, \gamma \geq 1$:

$$dx_1/dt = \alpha /(1 + z_1) - x_1, \quad z_1 = x_n^{\gamma} + x_{n-1}^{\gamma} + \ldots + x_{n-k+2}^{\gamma},$$
$$dx_2/dt = \alpha /(1 + z_2) - x_2, \quad z_2 = x_1^{\gamma} + x_n^{\gamma} + \ldots + x_{n-k+3}^{\gamma}, \qquad (4)$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$dx_n/dt = \alpha /(1 + z_n) - x_n, \quad z_n = x_{n-1}^{\gamma} + x_{n-2}^{\gamma} + \ldots + x_{n-k+1}^{\gamma},$$

where $1 < k \leq n$. The stable properties of these systems are described by the (n,k) – criterion (Likhoshvai et al., 2001). However, these systems have unstable boundary states too. Currently, we have a complete understanding about all the stationary points of the model M(n,k) (Fadeev et al., 2002). By numerical methods, included into the GeneNetSTEP complex, it is possible to extract important additional information about the properties of these systems.

In Figure, the plots are shown that illustrate dependency of the amplitudes of oscillations from α of three limiting cycles found in the model M(6,4). Out of these cycles, only two limiting cycles are stable. The first stable cycle has three major variables, $x_1(t)= x_3(t-2T_1/3)=x_5(t-T_1/3)$, with the maximal amplitudes 1, and three minor variables, $x_2(t)=x_4(t-2T_1/3)=x_6(t-T_1/3)$, with the maximal amplitudes 2. Analogously, the second stable cycle has three major variables, $x_2(t)=x_4(t-2T_1/3)=x_6(t-T_1/3)$, of the amplitude 1, and three minor variables, $x_2(t)=x_6(t-T_1/3)=x_4(t-2T_1/3)$, of the amplitude 2. T_1 is the period of stable cycles. The third limiting cycle was found to be unstable. It is characterized by the phase displacement of the variables relatively each other: $x_1(t)= x_2(t-T_2/6)= x_3(t-2T_2/6)= x_4(t-3T_2/6)= x_5(t-4T_2/6)= x_6(t-5T_2/6)$, and its amplitude

in dependence upon the value of the parameter α is determined by the curve 3. As seen in Fig.1, the cycles are stable, if $\alpha>3$, and unstable, if $\alpha>4$.

Acknowledgements

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

References

1. Kogai V.V., Fadeev S.I. (2001) Application of the parameter's extension method on the basis of the method of multiple shooting for numerical studying of nonlinear boundary problems. Sib. J. of Industrial Mathematics. 4, 83-101.
2. Likhoshvai V.A., Matushkin Yu.G. (2002) On the theory of prediction of global modes in the function of gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
3. Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2002) Study of the regimes of functioning of symmetric gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
4. Likhoshvai V.A., Matushkin Yu.G., Vatolin Yu.N., Bazhan S.I. (2000) A generalized chemical kinetic method for simulating complex biological systems. A computer model of λ phage ontogenesis. Computational Technologies. 5, P. 87-99.
5. Likhoshvai V.A., Matushkin Iu.G., Fadeev S.I. (2001) Relationship between a gene network graph and qualitative modes of its functioning. Mol Biol (Mosk). 35, 1080-1087.
6. Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. (2001b) A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk). 35(6), 1072-1079. (In Russ.).
7. Fadeev S.I., Klishevich M.A., Likhoshvai V.A. (2002) Qualitative and numerical studying of hypothetical gene networks by the example of the model M(n,n). Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
8. Fadeev S.I., Pokrovskaya S.A., Berezin A.Yu., Gainova I.A. (1998) The software package STEP for numerical studying of nonlinear systems of equations and autonomous systems of the general form. Description how to use the package STEP by the example of the learning tasks from the university course «Engineering chemistry of catalytic processes. Manual, NSU.
9. Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. (2000) Mathematical model of cholesterol biosynthesis regulation in the cell. Proc. of the second International conf. on bioinformatics or genome regulation and structure. Novosibirsk, 1, 199-202.
10. Pissanetsky S. (1988) Technology of sparse matrices. Mir, Moskva. (Russian).

QUALITATIVE AND NUMERICAL STUDYING OF HYPOTHETICAL GENE NETWORKS BY THE EXAMPLE OF THE M(N,N) MODEL

[1] Fadeev S.I., [2] Klishevich M.A., [3] Likhoshvai V.A.*

[1] Mathematical Institute of SB RAS, Novosibirsk, Russia, e-mail Fadeev@math.nsc.ru
[2] Novosibirsk State University, Novosibirsk, Russia
[3] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
[*]Corresponding author

Keywords: *hypothetical gene networks, mathematical model, regulation, negative feedback, positive feedback, critical points, limiting cycles, stability*

Resume

Motivation: Studying of the properties of theoretical constructions of gene networks is an essential stage for understanding regularities of gene network's functioning in nature.
Results: In this work, we have completely studied the stationary solutions of the autonomous system of equations representing the mathematical model M(n,n) of hypothetical gene networks. In particular, we have found all solutions, the total number of which equals to $2^n - 1$. The analysis of their stability has validated the (n,k) – criterion about existence of n asymptotically stable stationary points of the model M(n,n).

Introduction

For fundamental analyzing of regularities in gene network functioning, we have introduced a novel object, hypothetical gene networks (HGN) (Likhoshvai, Matushkin, 2002; Likhoshvai et al., 2001). The HGN consists of the units of two types: genetic elements and regulatory relations. HGNs are useful for studying the role of the positive and negative feedbacks in manifestation of the global properties of gene networks, beyond considering the individual peculiarities of structure and functioning of the gene networks in nature. In this work, we consider the properties of symmetric hypothetical gene networks of the class 1, the model M(n,k). For the M(n,k) model, we have suggested a hypothesis claiming that the number of the stationary stable solutions of the model, as well as that of limiting cycles, are determined in accordance with the rule named as the (n,k)-criterion (Likhoshvai et al., 2001; Likhoshvai et al., 2002). Following the (n,k)-criterion, the properties of the autonomous system with the parameters α and γ, representing the model M(n,k), under noticeable large α and γ, depend upon the greatest common devisor d, of the numbers n and k: the system has d stable stationary states, if d = k, and d stable limiting cycles, if d < k. All the other boundary states are unstable. As seen, the (n,k)-criterion produces a simple way for designing the model gene construction with the arbitrary number of stable stationary points or limiting cycles. The validation of the (n,k)-criterion that was deduced empirically is a problem of qualitative studying of the autonomous system (1).
In this work, we demonstrate that the model M(n,n) may have only symmetric or partially symmetric stationary points. The exhaustive treatment of the stationary points is given, including the description of their stability and calculation of their total number.

Implementation and Results

Symmetric HGNs are constructed in a following way. Two natural numbers are fixed, $2 \le k \le n$. Then we take n genetic elements and order them according to a cycle. Next, we suppose that each gene is an inhibitor of subsequent clockwise k-1 genes. For example, if k=2, then the first gene inhibits the second one, the second gene inhibits the third one and so on, the n-th gene inhibits activity of the first gene.
An example of symmetric HGN with six genetic elements (n=6, k=2), such that activity of each genetic element is suppressed by the product of synthesis encoded by genetic element with the lesser number is shown in Fig. 1.
The mathematical model M(n,k), of symmetric hypothetical gene networks of the class 1, which contain n genetic elements, is described by the autonomous system of n equations with the parameters $\alpha, \beta > 0, \gamma \ge 1$:

$$\begin{aligned}
&dx_1/dt = \alpha/(1+\beta z_1) - x_1,\ z_1 = x_n^{\gamma} + x_{n-1}^{\gamma} + \ldots + x_{n-k+2}^{\gamma},\\
&dx_2/dt = \alpha/(1+\beta z_2) - x_2,\quad z_2 = x_1^{\gamma} + x_n^{\gamma} + \ldots + x_{n-k+3}^{\gamma},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&dx_n/dt = \alpha/(1+\beta z_n) - x_n,\quad z_n = x_{n-1}^{\gamma} + x_{n-2}^{\gamma} + \ldots + x_{n-k+1}^{\gamma},
\end{aligned} \qquad (1)$$

where $1 < k \le n$. We are interested in answering the following questions: whether the stationary solutions of the system (1) do exist and which is the number of these solutions, whether the auto-oscillations are possible and how to make the stationary and periodical solutions.

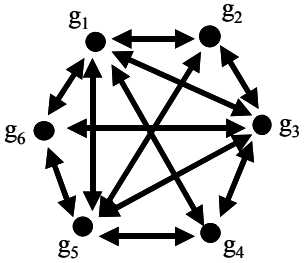


Fig. 1. An example of symmetric HGN. g1,…,g6 are genetic elements, arrows denote negative feedbacks, an arrow points to direction of the negative feedback.

Two properties of solutions of the system (1) are given below without proofs.

1. If the initial conditions of Cauchy problem for the system (1) are not negative and less than α, then solution of Cauchy problem exists for each $t > 0$, and this solution belongs to domain of positivity: $0 < x_i(t) < \alpha$, $i = 1,2,\ldots,n$.
2. The M(n,k) model may have no stationary solutions such that all components of them are different.

As a partially symmetric solutions, we denote such solutions of the system (1), in which some of the components $x_1, x_2, \ldots, x_n$ of the vector x coincide. The same assertions are referred to the stationary solutions determined by the system

$$f_i = \alpha/(1 + \beta z_i) - x_i = 0, \quad i = 1,2,\ldots,n. \tag{2}$$

Symmetric solutions of the model M(n,k)

By symmetric solution of the system (1), we denote a solution, in which all the components are equal:

$$x_1 = x_2 = \ldots = x_n = U_0. \tag{3}$$

One can readily see that symmetric stationary solution of the model M(n,k) is determined from the equation

$$\alpha = U_0[1 + \beta(k-1)U_0^{\gamma}]. \tag{4}$$

In the case given, analysis of stability of the symmetric stationary solution in dependence upon the parameter α is directly related to determining the spectrum of the Jacobi matrix, which is the circulant matrix. The proper numbers are calculated by the formulas given below:

$$\lambda_j = -1 - X\cdot(\varepsilon_j + \varepsilon_j^2 + \ldots + \varepsilon_j^{k-1}), \quad j = 1,2,\ldots,n, \tag{5}$$

where $\varepsilon_j = \cos(2\pi j/n) + i\cdot\sin(2\pi j/n)$, $X = \alpha\beta\gamma U_0^{\gamma-1}/[1 + \beta(k-1)*U_0^{\gamma}]^2$, $X \geq 0$.

Thus, at a point $X = 0$ (that is, $\alpha = 0$), symmetric stationary solution is asymptotically stable. Let d be the greatest common devisor of the numbers n and k. Then the spectrum of the circulant matrix (5), under $X = 1$, has the zero proper number with multiplicity $d - 1$. If $d=1$ (n and k are relatively prime numbers), then the circulant matrix is always non-degenerate. If n and k are not relatively prime numbers and $\gamma > k - 1$, then the circulant matrix is degenerated for

$$\alpha = \gamma U_0/(\gamma - k + 1), \; U_0 = 1/[\beta(\gamma - k + 1)]^{1/\gamma}. \tag{6}$$

Hence, for $\alpha > \gamma U_0/(\gamma - k + 1)$, the circulant matrix looses stability.

In case the circulant matrix is degenerated, (6) gives the value α, which determines essentially critical point of the system (2). If $d=1$, then we observe a Hopf bifurcation at this point.

Partially symmetric solutions of the model M(n,n)

The model M(n,n) is interesting, because an arbitrary partitioning of the integrity of components of the vector x into ν non-intersecting groups ($1 \leq \nu < n$) and equating of the components of the j-th group to the value U_j, $j = 1,2,\ldots,\nu$, generates the system of dimensionality $\nu < n$. Hence, all partially symmetric solutions of the initial system (2) could be found by analyzing the systems of lesser dimensionality.

The systems with $\nu = 2$ are of particular interest. In total, there exist [n/2] various types of systems, to which the initial system (2) could be reduced under n=k. If we denote as m ($m \leq [n/2]$) the number of components in the first group and equal them to U_1, whereas all the components of the second group equal to U_2, then the system (2) is rearranged to two equations relatively U_1 and U_2:

$$f_1 = \alpha/[1+\beta((m-1)U_1^{\gamma}+(n-m)*U_2^{\gamma})] - U_1 = 0,$$
$$f_2 = \alpha/[1+\beta(mU_1^{\gamma}+(n-m-1)U_2^{\gamma}] - U_2 = 0. \tag{7}$$

The system (7) assumes an exact solution in a form of parametric dependency of U_1, U_2, and α from the parameter $s = U_1/U_2$:

$$U_2(s) = [(1/\beta)/g(s)]^{1/\gamma}, \; U_1(s) = sU_2(s),$$
$$\alpha(s) = U_1(s)[1+\beta((m-1)U_1^{\gamma}(s)+(n-m)U_2^{\gamma}(s))], \tag{8}$$

where $g(s) = (s^{\gamma}-1)/(s-1)-(m-1)s^{\gamma}+m-n$, if $s \neq 1$, whereas $g(s) = \gamma + 1 - n$, if $s = 1$.

It is not difficult to reveal that symmetrical solution (i.e., $U_1=U_2$) and all partially symmetrical solutions of the system (7), as the function of the parameter α, intersect at the essentially critical point and that for every m and noticeably large values of α, that is, for $\alpha > \gamma U_0/(\gamma - n + 1)$, there exist two solutions of the system (7).

It turns out that solutions of (8) represent all stationary solutions of the model M(n,n). Let us represent the system (2), $k = n$, in a form:

$$x_j[1 + \beta(S - x_j^{\gamma})] - \alpha = 0,$$
$$S = x_1^{\gamma} + x_2^{\gamma} + \ldots + x_n^{\gamma}, \; j = 1,2,\ldots,n. \tag{9}$$

Let us consider an equation

$\alpha = u[1 + \beta(S - u^{\gamma})]$, (10)

where S is a parameter determining the function $u = u(S)$ under the set values of α, β, and γ.

The equation (10) orders the paraboloid-like dependency of α from u: $\alpha \geq 0$, in case $0 \leq u \leq u_s$, where $u_s = (1/\beta + S)^{1/\gamma}$. The maximal value of α, equaling to α_m , is achieved under $u = u_m$:

$\alpha_m = u_m\gamma(1 + \beta S)/(1 + \gamma)$, $u_m = [(1/\beta+S)/(1+\gamma)]^{1/\gamma}$.

For $0<\alpha<\alpha_m$, an equation (10) has two roots, $u_1(S)$ and $u_2(S)$, $0<u_1(S)<u_m$, $u_m<u_2(S)<u_s$. Turning back to the system (9), we see that every x_j equals either to $u_1(S)$, or to $u_2(S)$, thus, solution of the system (2), k = n, is always characterized by a partial symmetry, such that m components of the vector equal to $u_1(S)$, and n - m components equal to $u_2(S)$. The value S, determining solution of the system (1), is found from the equation:

$S = x_1(S)^{\gamma} + x_2(S)^{\gamma} + ...+ x_n(S)^{\gamma}$ (11)

As shown by direct calculations, for noticeably large α and $\gamma > n - 1$, the number of stationary solutions of the model M(n,n) equals to 2^n-1. Amongst them are symmetric solution and partially symmetric solutions with two groups of coinciding components. In Fig. 2, the plots are demonstrated, of dependence of the variable U_1 of the equations (7) and symmetric solution U_o upon the parameter α for n = 6, γ = 6. Thus, here are represented all the stationary solutions of the model M(6,6), the number of these solutions under $\alpha>6$ being equal to 2^6-1=63.

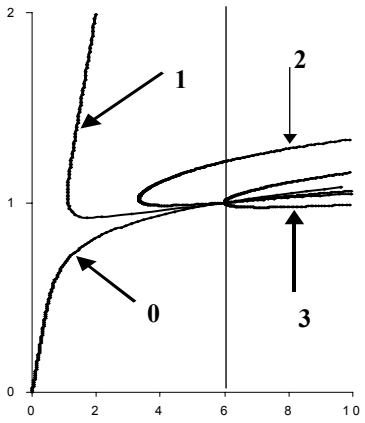


Fig. 2. Dependencies of the variable U_1 of equations (7) and symmetric solution U_o from the parameter α for n = 6, γ = 6. 0, symmetric solution; 1, U_1 corresponds to m=1; 2, U_2 corresponds to m=2; 3, U_2 corresponds to m=3.

Numerical studying of stability made by the software package STEP (Fadeev et al., 2002) have shown that only n stationary solutions are asymptotically stable, these solutions corresponding to the only branch of solution of the system (7), for m = 1. That is, one of the components of the vector x equals to U_1, while the rest components equal to U_2. Thus, the (n,k)-criterion for the model M(6,6) is completely verified.

Acknowledgements

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

References

1. Likhoshvai V.A., Matushkin Yu.G. (2002) On the theory of prediction of global modes in the function of gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002), this issue.
2. Likhoshvai V.A., Matushkin Iu.G., Fadeev S.I. (2001) Relationship between a gene network graph and qualitative modes of its functioning. Mol. Biol. (Mosk). 35, 1080-1087.
3. Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2002) Studying of the regimes of functioning of symmetric gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002), this issue.
4. Fadeev S.I., Berezin A.Yu., Gainova I.A., Kogai V.V., Ratushny A.V., Likhoshvai V.A. (2002) Development of the program software for mathematic modelling of the gene network dynamics. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002), this issue.

DETERMINATION OF BIFURCATIONAL PARAMETER VALUES OF MATHEMATICAL MODEL **M(n,k)** OF HYPOTHETICAL GENE NETWORKS

[1] Fadeev S.I., [2] Vernikovskaya E.V., [2] Purtov A.V., [3] Likhoshvai V.A.*

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

Keywords: *hypothetical gene networks, mathematical model, regulation, negative feedback, positive feedback, critical points, limiting cycles, stability, bifurcation values*

Resume

Motivation: Studying of the phase-plane portraits of mathematical models of hypothetical gene networks is an essential stage in understanding the regularities of gene network's functioning in nature.
Results: In this work, bifurcation parameter values of the mathematical model M(n,k) are calculated. The limiting cycles of these models are being analyzed. The simple way of searching for the limiting cycles by solving the boundary case of the equation with retardation is presented.

Introduction

For making comprehensive analysis of regularities in the gene network functioning, we have introduced the theoretical object, a hypothetical gene network (HGN) (Likhoshvai, Matushkin, 2002; Likhoshvai et al., 2001). The HGNs are built out of the units of two types: genetic elements and regulatory relations. If to abstract from the particular individual properties of a gene network's structure and functioning in nature, then the HGNs could be useful for studying the role of the negative and positive feedbacks in formation of the global properties of a gene network. In this work, we consider symmetric hypothetical gene networks of the class 1, which are described by the models M(n,k) (denotation of classes is given elsewhere (Likhoshvai, Matushkin, 2002)). We analyze the critical values of the parameters of the model M(n,k), which are responsible for alteration of properties in the model's limiting behavior. Application of the methods of qualitative and numerical analyses gave the detailed representation about the properties of the model, thus, supplementing the assertions of the (n,k) – criterion.

Implementation and Results

The symmetric HGNs are constructed as given below. Two natural numbers, $2 \le k \le n$, are fixed. Then we take n genetic elements and order them according to a cycle. Next, we suppose that each gene is an inhibitor of the subsequent k-1 genes located clockwise. For example, if k=2, then the first gene inhibits the second one, the second gene inhibits the third one and so on, the n-th gene inhibits activity of the first gene. An example of symmetric HGN with six genetic elements (n=6, k=2), such that activity of each genetic element is suppressed by the product encoded by genetic element with the lesser number is shown in Fig. A mathematical model M(n,k) of symmetric hypothetical gene network of the class 1, which consists of n genetic elements is described by the autonomous system of n equations with the parameters $\alpha, \beta > 0, \gamma \ge 1$.

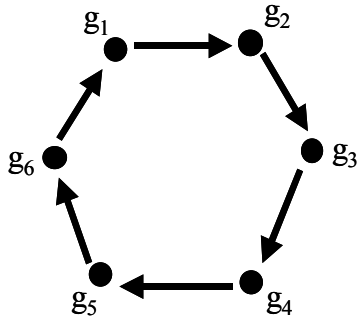


Fig. An example of symmetric HGN. g1,…,g6 are genetic elements, arrows denote negative feedbacks, an arrow points to direction of the negative feedback.

A symmetric HGN of the class 1 is described by the model M(n,k) of the form:

$$dx_1/dt = \alpha/(1 + z_1) - x_1, \quad z_1 = x_n^{\gamma} + x_{n-1}^{\gamma} + \ldots + x_{n-k+2}^{\gamma},$$
$$dx_2/dt = \alpha/(1 + z_2) - x_2, \quad z_2 = x_1^{\gamma} + x_n^{\gamma} + \ldots + x_{n-k+3}^{\gamma}, \qquad (1)$$

. . . . . .

$dx_n/dt = \alpha/(1 + z_n) - x_n, \quad z_n = x_{n-1}^{\gamma} + x_{n-2}^{\gamma} + \ldots + x_{n-k+1}^{\gamma},$

where $\alpha > 0$, $\beta > 0$, $\gamma \geq 1$ are the parameters, $1 < k \leq n$. As a result of searching for dependency of the stationary solutions of (1) from the parameters α and β, we have determined their bifurcation values, that is, the values, when the number of stationary solutions changes or the pattern of stability alters, including the case when stability of the stationary solution is lost and the stable limiting cycle is formed (Hopf bifurcation). Parametric analysis is mainly based on existence of exact representation of solutions of the system $f_i = \alpha/(1 + \beta z_i) - x_i = 0, \quad i = 1,2,\ldots,n,$ (2) which determines the stationary solutions of (1).

Symmetric Solution

The model M(n,k) has a symmetric solution, that is, the solution such that all its components coincide:
$x_1 = x_2 = \ldots = x_n = U_0$,
where dependency of U_o from α is determined from the equation:

$$\alpha = U_0[1 + \beta(k-1)U_0^{\gamma}]. \qquad (2)$$

An important is the fact that the Jacobi matrix of the system (2) for symmetrical solution is a circulant matrix and, hence, the proper numbers of the circulant matrix have an exact expression. In this case, we may write in an explicit form the dependency of the proper numbers from the parameter α.

If n and k are not relatively prime numbers with the greatest common divisor $d > 1$, then there exists a value $\alpha=\alpha_o$, such that the pattern of stability of symmetric solution changes: for $\alpha<\alpha_0$, symmetric solution is asymptotically stable, whereas for $\alpha>\alpha_0$, unstable. The bifurcation value α is calculated by the formula:

$$\alpha_0 = \gamma U_0/(\gamma - k + 1), \quad U_0 = 1/[\beta(\gamma - k + 1)]^{1/\gamma}, \ \gamma > k - 1. \qquad (3)$$

As a consequence, there exists a bifurcation value $\gamma_0 = k - 1$: for $\gamma < \gamma_0$, symmetric solution is asymptotically stable for each $\alpha > 0$.

If n and k are relatively prime numbers, then the existence of α_0 and γ_0 is directly related to the outlet on the imaginary axis of the pair of complex-conjugate proper numbers (Hopf bifurcation). Since for the model M(n,2), n is an odd number, then

$\alpha_0 = U_0/(1 - \gamma_0/\gamma)$, $U_0 = 1/[\beta(\gamma/\gamma_0 - 1)]^{1/\gamma}$, $\gamma > \gamma_0$, where $\gamma_0 = \cos(\pi/n)$.

Partially symmetric solutions

For $d > 1$, except the symmetric solution, the system (2) have partially symmetric solutions, in which components x_i, $i = 1,2,\ldots,n$, are subdivided into two groups. In each group, the component values coincide and equal to U_1 and U_2, respectively. Then the system (2) could be rearranged into the system of two equations relatively U_1 and U_2. If $d = 1$, then the system (2) has only the symmetric solution, whereas the stationary solutions of the other type for the model M(n,k) are absent. This fact is proved for $k = n$ and $k = 2$ and it is verified numerically for the intermediate values of k.

In accordance with the (n,k) – criterion, three cases are possible.

1) Let $d = 1$. Then for $\alpha = \alpha_0$, there exists a Hopf bifurcation. For $\alpha > \alpha_0$, the only solution of the system (2) becomes unstable and, hence, auto-oscillations are generated.

2) If $1 < d < k$, then $\alpha = \alpha_0$ is essentially a critical point of the system (2), which determines the location of intersection of the plots illustrating symmetrical and partially symmetrical solutions. For $\alpha>\alpha_0$, all partially symmetrical and symmetrical solutions are unstable, thus, auto-oscillations are generated.

3) If $d = k$, then the stable limiting cycles are absent. For $\alpha = \alpha_0$, an essentially critical point determines the point of intersection of the plots illustrating symmetrical and partially symmetrical solutions, among which k solutions are stable, if $\alpha > \alpha_0$.

Note that the limiting cycles of the model M(n,k) are also characterized by a symmetry, or by a partial symmetry. In case of symmetry, all components of solution coincide by an amplitude and differ only by the phase displacement. In case of partial symmetry, as in the case of stationary solutions, we have two groups of coinciding by an amplitude components that differ from each other only by the phase displacement.

Critical points of the type "turning point"

Let us pay attention to numerical determination of bifurcation values of the parameters of the model M(n,k). Application of the software package STEP (Fadeev et al., 2002), enables to make effectively the numerical study of solutions of the model M(n,k), in particular, to study dependence of stationary solutions upon the parameter α according to the method of continuation by a parameter, supplemented by analysis of stability. In this case, the formulas providing the exact solution under some value of the parameter α, enable to order the starting solution. Simultaneously to constructing dependency of solution from the parameter α, we search for bifurcation values of α, which are the turning points, on the plot, of some partially symmetric solutions. Hence, together with the essentially critical point, these data could be used for searching for the ranges of variation of the parameter α with different number of stationary solutions.

By the same method of continuation by a parameter (PCM), with the help of the software package BPR-Q, it is possible to study numerically dependency of the limiting cycles upon the parameter α as dependency on α of solution of the nonlinear

boundary problem (Kogai, Fadeev, 2001). In this case, the critical points referring to the type "turning point" could be also found. In the cases, when the limiting cycles were found to be multiple, the correspondence to the (n,k) – criterion is detected only after clearing up their stability. All the "extra" limiting cycles, from the point of view of the (n,k)-criterion, appear to be unstable.

On unstable boundary solutions of the model M(n,k)

As was noted above, the stationary partially symmetric solutions of the model M(n,k) satisfy formally to the system of two equations relatively U_1 and U_2. If the initial conditions of Cauchy problem will belong to variety, determined by the form of partial symmetry, then Cauchy problem's solution will also fall to this variety.

Studying of solutions for varieties, that is, solutions of Cauchy problems, formulated relatively U_1 and U_2, enables to consider unstable stationary solutions, or unstable limiting cycles of the model M(n,k), with their subsequent analysis according to the PCM and searching for the critical points of the type "turning point".

Usage of partial symmetry of solutions representing the limiting cycles of the system (1) reveals one more approach to studying the limiting cycles, for which the problem is formulated as a boundary problem for one or two differential equation with retarded argument. For example, in case of the model M(n,2), where n is an odd number, the study of symmetric auto-oscillations in dependence upon the parameter α could be reduced to the following boundary problem:

$$du/dt = T(\alpha/[1 + \beta u^{\gamma}(t - (1-1/n)/2)] - u), \quad (4)$$

$$t \in [0,1],\ u(0) = u(1),\ \alpha = u(0)[1 + \beta u^{\gamma}(-(1-1/n)/2)].$$

Here T is a period of auto-oscillations that should be determined, the boundary conditions express the periodicity and transversality conditions. Different ways could be mentioned for constructing the discrete model of the boundary problem (4), which gives an approximate representation of the problem in a form of the system of nonlinear equations relatively the values of the sought function u(t) in the mesh points with subsequent studying of the system by the PCM and searching for the critical points of the type "turning point". Here, accounting for retardation is expressed due to the fact that from condition of periodicity u(t) follows the equality: $0 \le t \le (1-1/n)/2$, $u(t) = u(t + (1+1/n)/2)$.

Again, we draw an attention to undoubtedly interesting and methodically perspective fact: the study of symmetric limiting cycles for M(n,2) is reduced to analysis of a single equation. Application of this observation to the models M(n,2) enables to detect in them a considerable number of unstable partially symmetric limiting cycles. The mechanism for the searching for such cycles is extremely simple. One should take some factorization n into efficients q and p (p is an odd number) and consider the system (4) for n=p. The resulting limiting cycle of the system (4) will be unstable cycle of the system (1), which is stable for variety $x_i = x_{i+p} = \ldots = x_{i+(q-1)p}$, $i = \overline{1, p}$.

Acknowledgements

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

References

1. Likhoshvai V.A., Matushkin Yu.G. (2002) On the theory of prediction of global modes in the function of gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
2. Likhoshvai V.A., Matushkin Iu.G., Fadeev S.I. (2001) Relationship between a gene network graph and qualitative modes of its functioning. Mol Biol (Mosk). 35, 1080-1087.
3. Kogai V.V., Fadeev S.I. (2001) Application of the parameter's extension method on the basis of the method of multiple shooting for numerical studying of nonlinear boundary problems. Sib. J. of Industrial Mathematics. 4, 83-101.
4. Fadeev S.I., Berezin A.Yu., Gainova I.A., Kogai V.V., Ratushny A.V., Likhoshvai V.A. (2002) Development of the program software for mathematic modelling of the gene network dynamics. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

ANALYSIS OF PROPERTIES OF HYPOTHETICAL GENE NETWORKS WITH POSITIVE FEEDBACK

[1] Fadeev S.I., [2] Osokina V.A., [3] Likhoshvai V.A.

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

Key words: *hypothetical gene networks, mathematical model, regulation, negative feedback, positive feedback, critical points, limiting cycles, stability, equivalence of limiting properties for the systems of differential equations*

Resume

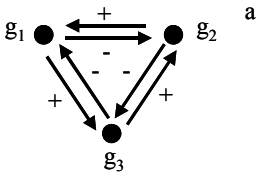
Motivation: A necessary stage in understanding of a gene network function in nature is to determine the impact of positive and negative feedbacks on theoretical constructions of gene networks.
Results: In this work, we have studied the properties of symmetric hypothetical gene network that consists of three genetic elements, three negative feedbacks, and three positive ones. As demonstrated, the limiting properties of this network are equivalent to the properties of two hypothetical gene networks without positive feedbacks.

Introduction

In the functioning of a gene network, the primarily, if not determinative role, belongs to genes encoding the regulatory proteins that are capable to activate and suppress the action of other genes. Due to the action of these proteins, the gene networks gain ability to auto-regulation and proper responsiveness to alterations of external environmental conditions (Kolchanov, 1997). In this connection, it is of pivotal importance to study the role of the negative and positive feedbacks in a gene network functioning.
One of the approaches is to use the mathematical modelling of theoretical constructions of gene networks, in which the relations between positive and negative feedbacks could be studied by disengaging from other details of the gene network structure. Previously, we have studied the properties of hypothetical gene networks (HGNs) with negative feedbacks. As was demonstrated, the stable states of these networks (stationary points, continuous waves) in particular parametric intervals are determined only by the structure of relations between the genes (Likhoshvai et al., 2001). In this work, we study the properties of symmetric hypothetical gene network consisting of three genetic elements, three negative and three positive feedbacks (Fig. 1a). We have revealed that for the gene network considered there are two parametric intervals. In the first interval, there are three stable points, that is, the properties of this gene network are similar to that of the network, where the products of each three elements are inhibitors of activity of another two genetic elements (Fig. 2a). In the second parametric interval, the gene network is characterized by a single continuous cycle, similar to symmetric hypothetical gene network, in which each gene product is an inhibitor of activity of the next-in-turn genetic element (Fig. 3a). As follows from the results obtained, the positive feedbacks do not make a novel impact in a functioning of gene networks, in which all regulatory relations are negative. This points to the principal possibility to reduce description of more complex gene networks to considering the properties of the limited number of basic gene networks with more simple construction.

Implementation and Results

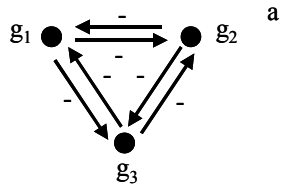
One of the approaches of studying the properties of hypothetical gene networks is estimation of equivalence of the HGN's model behavior in a definite parametric interval corresponding to behavior of the basal model of the hypothetical gene network with well-studied properties. To basal HGN, we may refer, for example, the HGNs described by the mathematical model M(n,k) (Likhoshvai et al., 2001). Some integrity of the limiting conditions of this model, which is the system of differential equations, we consider as equivalent to the integrity of the limiting conditions of the other model, if there exists continuous transformation of one model into another, such that each condition of the first integrity is transformed continuously into some condition of the second integrity and *vice versa*. The methods like the parameter continuation method (PCM) (Fadeev et al., 2002a) enable to reveal successfully dimensions of the intervals of equivalence.



$$dx_1/dt=(\alpha_0+\alpha_1 x_2^{\gamma})/(1+\beta(x_2^{\gamma}+x_3^{\gamma}))-x_1$$
$$dx_2/dt=(\alpha_0+\alpha_1 x_3^{\gamma})/(1+\beta(x_1^{\gamma}+x_3^{\gamma}))-x_2$$
$$dx_3/dt=(\alpha_0+\alpha_1 x_1^{\gamma})/(1+\beta(x_1^{\gamma}+x_2^{\gamma}))-x_3$$

Fig. 1. (a) Structural graph of HGN with the positive and negative feedbacks. (b) Mathematical model $MX_1(3,3)$, α_0, α_1, β, and γ are positive parameters. For details, see the text.

Let us consider the HGN with the structural graph given in Fig. 1a. The vertexes of the graph are represented by genetic elements. Thus, the HGN contains three genetic elements. The edges oriented clockwise and marked by "-" correspond to the negative feedbacks. The edges oriented anticlockwise (marked as "+") correspond to the positive feedbacks. The mathematical model describing this HGN is illustrated in Fig.1b. In what follows, we shall denote it as the model $MX_1(3,3)$. Following classification given in (Likhoshvai, Matushkin, 2002), this HGN is referred to the class 1.



$$dx_1/dt=\alpha/(1+\beta(x_2^{\gamma}+x_3^{\gamma}))-x_1$$
$$dx_2/dt=\alpha/(1+\beta(x_1^{\gamma}+x_3^{\gamma}))-x_2$$
$$dx_3/dt=\alpha/(1+\beta(x_1^{\gamma}+x_2^{\gamma}))-x_3$$

Fig. 2. (a) Structural graph of the complete symmetric HGN with negative feedbacks. (b) Mathematical model M(3,3), α, β, and γ are positive parameters. For the comments, see the text.

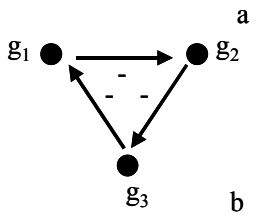
The main goal of this work is to compare the model $MX_1(3,3)$ to the models M(3,2) (Fig. 2b) and M(3,3) (Fig. 3b), which describe the HGNs illustrated in Fig. 2a and Fig.3a, respectively. In the models M(3,3) and M(3,2), the positive feedbacks are absent.

Following the (n,k)-criterion, for sufficiently large α and γ, the model M(3,3) has three stable stationary solutions (Likhoshvai et al., 2002). More detailed analysis gives evidence that the total number of stationary solutions, including the unstable ones, equals to 7 (Fadeev et al., 2002). Among these solutions are the symmetrical one, $x1=x_2=x_3=U_0$, and 6 partially symmetrical solutions with the partial symmetry of the form: $x_1 = U_1$, $x_2 = x_3 = U_2$, or $x_2 = U_1$, $x_1 = x_3 = U_2$, or $x_3 = U_1$, $x_1 = x_2 = U_2$.

Notably, there exist the partially symmetrical solutions, including the stable ones entering the critical point of the system of equations for determination of the stationary solutions of the model M(3,3), for the values $\alpha > \alpha^*$ and $\gamma > 2$, where α^* is determined as the turning point of the partially symmetrical solution. Symmetrical solution becomes unstable to the right of essentially critical point.

For the model $MX_1(3,3)$, let us set $\alpha_1=0$ and take a value of α_0 such that symmetric solution of the model M(3,3), for $\alpha=\alpha_0$, is unstable. Then there exists some neighborhoods of the point $\alpha_1= 0$, where the main properties of the model $MX_1(3,3)$ are conserved. By the continuation by the parameter α_1, it is possible to construct the diagram of the stationary solutions of the model $MX_1(3,3)$. In Fig. 4a, one may see the plots illustrating alteration of the stationary values of variables of the model $MX_1(3,3)$ in dependency upon the value of the parameter α_1, calculated for $\alpha_0=3$, $\beta = 1$, $\gamma = 4$. Here the curve u0 is a plot of symmetrical solution of the model $MX_1(3,3)$, which in the neighborhoods of $\alpha_1 = 0$ is unstable, due to the choice of α_0, while under $\alpha_1 > 0.66$, it becomes stable (Fig. 4a, vertical curve 1). At the boundary of stability of symmetrical solution, one may observe the Hopf bifurcation. The other branch appears by PCM from the partially symmetrical solution of the model $MX_1(3,3)$ (Fig. 4a, curves x1, x2, x3). The upper parts of the branches x1, x2, and x3, up to the turning point, $\alpha_1=0.708$ (Fig. 4a, vertical line 2), represent the stable stationary solution.

It is important to note that in the case considered, the turning point is located to the right of the point, where stability of symmetrical solution is lost, hence, under the parameters given in the model $MX_1(3,3)$, the interval with self-generation of auto-oscillations is absent. To the right of the turning point, the model $MX_1(3,3)$ has a single symmetric solution, which is stable. In the interval $0.66<\alpha_1<0.708$, the system has four stable points, whereas for $\alpha_1<0.66$, there are three stable stationary points, as in case of the model M(3,3).



$$dx_1/dt=\alpha/(1+\beta x_3^{\gamma})-x_1$$
$$dx_2/dt=\alpha/(1+\beta x_1^{\gamma})-x_2$$
$$dx_3/dt=\alpha/(1+\beta x_2^{\gamma})-x_3$$

Fig. 3. (a) Structural graph of symmetric HGN with three negative feedbacks. (b) Mathematical model M(3,2), α, β, and γ are positive parameters. See also comments in the text.

With the growth of α_0, the turning point moves relatively the point of loosing the stability of symmetrical solution and, at last, it becomes located to the left of this point. This situation, realized for $\alpha_0=7$, $\beta=1$, $\gamma=4$, is shown in Fig. 4b (curve 2 is situated to the left of the curve 1). Here, in the interval of variation of the parameter α_1, between the turning point and the point of loosing the stability of symmetrical solution, there exist a single symmetrical solution, which is unstable. Hence, in this interval, there is a generation of auto-oscillations. To the right of the curve 1, the only symmetrical solution is stable. To the left of the curve 2, there are three stable stationary points. In Fig. 5, one can see an example of the way out to auto-oscillations, of Cauchy problem solution, for the model $MX_1(3,3)$ under $\alpha_0=7$, $\alpha_1=1.1$, $\beta=1$, $\gamma=4$.

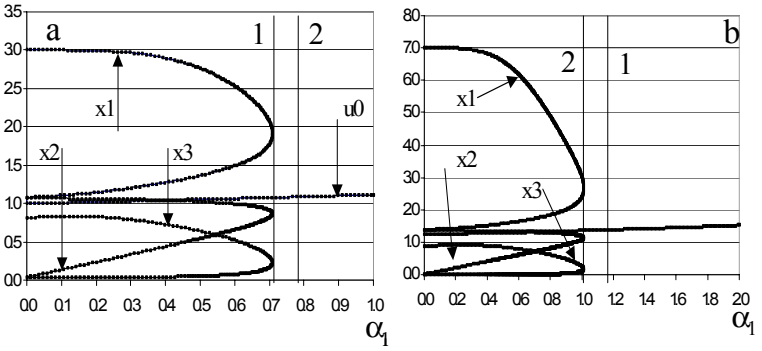


Fig. 4. Alteration of stationary values of the variables of the model $MX_1(3,3)$ in dependence upon the parameter α_1. By ordinate, the conditional value of the variable is shown. For explanations, see the text.

Let us consider the model M(3,2). By applying the homotopy method, by changing the parameter α_1 for the parameter $\alpha_1(1-q)$, let us "load" the model M(3,2) into the model $MX_1(3,3)$. Obviously, for $q=0$, we have the model $MX_1(n,k)$, whereas for $q=1$, the model M(3,2). By applying the method of the parameter extension of solution for the parameter q from 0 to 1 in case of the model $MX_1(3,3)$, where $\alpha_0=7$, $\alpha_1=1.1$, $\beta=1$, $\gamma=4$, we continuously, without the loss of limiting properties (the limiting cycle is conserved for every value of the parameter $q\in[0,1]$), pass from the model $MX_1(3,3)$ to the model M(3,2) with the known properties. That is to say, the model M(3,2) has a single stationary solution, while stability is lost under $\alpha=\alpha_0>3$. Thus, by parameter continuation, we estimate the equivalence of the stable phase-plane portrait of the model $MX_1(3,3)$, with the parameters $\alpha_0=7$, $\alpha_1=1.1$, $\beta=1$, $\gamma=4$, to the stable phase-plane portrait of the model M(3,2), with $\alpha=7$, $\beta=1$, $\gamma=4$. Analogously, it could be demonstrated that for $\alpha_0=7$, $\alpha_1=0.5$, $\beta=1$, $\gamma=4$, three available stable conditions of the model $MX_1(3,3)$ are equivalent to three stable conditions of the model M(3,3) (calculations are not shown).

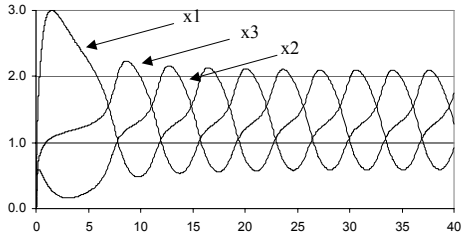


Fig. 5. A way out of Cauchy problem solution to auto-oscillations for the model $MX_1(3,3)$, under $\alpha_0=7$, $\alpha_1=1.1$, $\beta=1$, $\gamma=4$, and initial values x1=0.5, x2=x3=0. By abscissa, the conditional time; by ordinate, the conditional values of variables.

Conclusion

Thus, we have made a study of properties of a symmetric hypothetical gene network that consists of three genetic elements, three positive feedbacks and three negative ones (Fig. 1). We have shown that this network has two parametric intervals. Within the first interval, the properties of the network are similar to that of hypothetical gene network, in which the product of each of its genetic elements is an inhibitor of two other genetic elements (Fig. 2). Within the second interval, the stable behavior of this network is similar to that of symmetric hypothetical gene network, in which each product is an inhibitor of the consequent genetic element (Fig. 3). This result gives evidence that positive feedbacks do not supply the gene network with novel properties, which were not typical for a gene network possessing only by negative feedbacks. This study points to the principle difference between the roles of positive and negative feedbacks in formation of the properties of gene networks: the main role belongs to the stabilizing negative feedbacks, which are responsible for generation of the whole variety of properties of the gene networks, whereas the positive feedbacks only redistribute these properties over various parametric intervals.

As follows from this work, in principle, it is possible to reduce the complex gene networks to studying more simple basic gene networks, including HGNs described by the M(n,k)-models: for q = 0, we arrive at the model $MX_1(3,3)$, whereas for q=1, the model M(3,2).

Acknowledgements

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

References

1. Kolchanov N.A. (1997). Regulation of transcription of genes in eukaryotes: databases and computer analysis. Mol. Biol. (Mosk.). 31, 581–583.
2. Likhoshvai V.A., Matushkin Yu.G. (2002) On the theory of prediction of global modes in the function of gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
3. Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2001) Relationship between a gene network graph and qualitative modes of its functioning. Mol. Biol. (Mosk.). 35, 1080-1087.
4. Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. (2002) Studying the regimes of functioning of symmetric gene networks. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
5. Fadeev S.I., Berezin A.Yu., Gainova I.A., Kogai V.V., Ratushny A.V., Likhoshvai V.A. (2002a) Development of the program software for mathematical modelling of the gene network dynamics. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).
6. Fadeev S.I., Klishevich M.A., Likhoshvai V.A. (2002b) Qualitative and numerical studying of hypothetical gene networks by the example of the M(n,n) model. Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002).

A NOVEL ALGORITHM FOR *IN-SILICO* EST EXPRESSION PROFILING

**** Leyfer D.*** [1]***, Funari V.*** [1]***, Berwick R., Haverty P., Frith M., Tolan D.***

Boston University, Boston, USA, e-mail: dmitriyl@bu.edu
[*] Presenting author
[1] Both authors contributed equally to this project

Key words: *gene expression, expression profiling, transcriptional profiling, EST, transcriptome, dbEST, biological databases, in-silico, cDNA libraries, genomics, functional genomics*

Resume

Motivation: Various tasks in computational biology, including primer and oligonucleotide array design; epitope mapping, etc. require finding unique regions in gene sequences. We developed a novel algorithm that finds such unique regions based on an alignment of the gene sequence to its paralogs. This algorithm was utilized for *in-silico* expression profiling using EST databases – an accurate low cost alternative to high-throughput "wet bench" expression profiling methods.
Results: Analysis of expression patterns of enzymes in glycolytic pathway led to alternative hypothesis of fructose metabolism in the brain. Publicly available database of Expressed Sequence Tags (dbEST) consisting of multiple flat files was mined for crucial information and reformatted into a relational database. A public Internet access to the database is planned.
Availability:
-Available as a commercial package through Boston University technology transfer.
-Free availability over the Internet: watch for update at http://zlab.bu.edu/~dmitriyl.

Introduction

With the completion of the human genome sequencing the focus of biological research has shifted to 'postgenomics': gene expression analysis, signal transduction pathways, modelling biological processes. Various high-throughput methods of expression profiling are now common, but costly and labor-intensive. A low cost, fast alternative to wet bench transcriptional profiling is utilizing information from public and private EST databases.
ESTs (Adams et al., 1991) are single-pass sequenced cDNAs representing expressed genes from a specific cell population. EST library is a collection of ESTs from a single experiment. Database of Expressed Sequence Tags (dbEST, http://www.ncbi.nlm.nih.gov/dbEST/) is a public domain collection of flat files containing information about ESTs. Current number of ESTs in dbEST is close to 9 million and growing exponentially, the number of EST libraries exceeds 8000 and the number of species is 345. Large EST databases have also been compiled in some of the genomics companies. If one compares an EST library to a single gene array type, the benefits of complementing other high throughput expression profiling technologies with EST data become obvious.
Gleaning reliable expression information from EST data is challenging. One of the problems is that commonly used algorithms are designed not for EST expression profiling, but for assembling ESTs into contigs (clusters) for resolving full length cDNAs of novel genes. A by-product of such clustering is a collection of gene-specific ESTs, from which expression information can be derived. Such "misapplication" of algorithm could result in misinformation and errors. High sequence error rate (3.3%), alternatively spliced genes, 2-pass (instead of single pass) sequenced ESTs and contamination of dbEST with vector and wrongly indicated species sequences add to the challenge. Another common problem is the libraries in which the ratio between highly expressed and low abundance genes was altered to find rare transcripts. Such libraries are not suitable for quantitative analysis. If only quantitative libraries are used, then the number of ESTs for a certain gene can quantitatively represent the expression level of this gene in a tissue from which the ESTs originated (Funari et al., 2000). Yet another problem is inconsistent dbEST annotation that complicates data mining.

Algorithms

Virtual Northern Blot. While many of the EST-clustering algorithms are EST-centric, i.e. contig is assembled by "walking" from one EST to another, our approach is gene (cDNA)-centric. VNB starts with a gene sequence that is computationally divided, based on an alignment of this gene to its paralogs, in multiple probes that are unique for this gene. These probes are then computationally "hybridized" with identical sequences in dbEST to find ESTs corresponding to the gene. The number and the length of the probes are optimized based on several parameters that allows for high sensitivity and specificity. Using 100% sequence identity in the probe-EST alignment makes virtually certain that EST is specific for the gene, which eliminates the need for choosing an arbitrary cut-off, which is a major problem with identifying ESTs using BLAST, another gene-centric approach (Peri et al., 2001).

AutoProbe. AutoProbe is the core of VNB; it is an algorithm that finds maximally unique regions in a gene sequence based on a multiple alignment of the gene's cDNA and its paralogs. AutoProbe idea is taken directly from experimental molecular biology, where one uses a short nucleic acid probe to hybridize to complementary mRNA sequences in Northern Blot. Unlike in regular Northern blot, multiple probes along the entire cDNA length have to be used in VNB in order to pull down all the ESTs for this cDNA, given that the ESTs can correspond to any cDNA region.
The probes for virtual hybridization have to satisfy the following criteria:
1) Probe length has to be long enough in order NOT to pull random sequences from the database.
2) Probe length has to take into account the EST error rate.
3) The probes have to be maximally gene-specific, i.e. have the least similarity to the paralogs.
The minimum length requirement (1) is satisfied by applying Erdos-Renyi Law (Erdos, R'enyi, 1970). The maximum length (L) of a random sequence in the database of length (D) is defined by

$$L=\log_{1/P}(DM/\square) \quad (1)$$

where M is the number of probes, P is the probability of encountering any one of the nucleotides = 1/4 and α is a desired significance level. Although the total length of human dbEST is 2 GB (2 billion nucleotides), the non-redundant portion of it is only ca. 100 million nucleotides – the overall length of the coding regions. The number of probes for an average cDNA and the window size of 10 is 250. Substituting for D, M and α we get L = 19.43 for α= 0.05 and 20.59 for α=0.01, i.e. 20 nucleotides should be the lower limit of probe size in order not to extract random sequences from the database. This lower limit varies slightly with increasing probe length and the size of the database in case of organisms with higher than human length of coding regions.
The requirement (2) determines the upper size of the probe: an average EST error rate is 3.3%, therefore one can expect a sequencing error every 30 nucleotides. In order for a probe to be on average between sequencing errors, the probe should be no longer than 30 nucleotides. To satisfy gene-specificity requirement (3), each probe is given a score that reflects the similarity of the probe region to the paralogs. The score for a probe is calculated as follows: each nucleotide position in the multiple alignment is given a numerical value based on the number of matches, mismatches and gaps. The scores for each individual position are summed across the probe length. Matches are given higher score than mismatches while mismatches are higher than gaps, ensuring that the regions of the least similarity have the lowest score. In order to cover the entire length of the cDNA, each probe is chosen inside a fixed length sliding window based on the minimum probe score in this window. The length of the sliding window as well as the length of the probe could vary for each gene family depending on the sequence similarity between family members, and could be determined experimentally. We showed that for Aldolase C sliding window of 12 with probe size of 24 nucleotides are the parameters achieving the most sensitivity at 100% specificity.

Results

VNB was used to study expression profiles for fructose metabolism specific enzymes: aldolase isozymes and ketohexokinase (KHK). The goal was to obtain information about possible alternative sites of fructose metabolism, important for our understanding of Hereditary Fructose Intolerance (HFI) – a metabolic disorder in which unassimilated fructose-1-phosphate accumulates in liver, eventually shutting down gluconeogenesis and glycogenolysis, resulting in severe hypoglycemia, hepatic failure and eventually death. Although HFI patients have a mutation in one of the essential enzymes in the fructose metabolism, Aldolase B, a fraction of consumed fructose (ca. 40%) is processed by unclear mechanisms. The liver and kidneys, and to a lesser extent the small intestine, were the only organs reported to carry out this process, however in normal metabolism only ca. 60% of fructose is known to be internalized in these organs. Our strategy was to find alternative metabolic sites by identifying tissues that coexpress pathway-specific enzymes. Expression of KHK has been previously assayed by several different techniques, however the results were inconclusive (Table).

Table. KHK expression in brain.

Method	Result
Immunohistochemistry (Bergbauer et al., 1996)	-
Activity assays (Aldeman et al., 1967, Bais et al., 1985)	-/+
RNase Protection Assay (Hayward et al., 1998)	-
Affymetrix GeneChips (Haverty, 2001)	-
RT-PCR (Hayward et al., 1998)	+/-

VNB demonstrated that KHK exhibits a previously unknown expression in brain, colon and mammary gland. To confirm VNB results we performed RNA in-situ hybridization (RISH, data not shown) on brain sections with digoxigenin-labeled KHK antisense probe that confirmed KHK expression in cerebellum and brain stem. These results suggest certain regions of brain as alternative sites of fructose metabolism. Aldolase C, not Aldolase B isozyme was found to be coexpressed with KHK in brain, suggesting that alternative metabolic sites might use alternative isozymes in the same pathway. These hypotheses have yet to be confirmed by other methods, for example, activity assays on specific tissues/primary cell cultures or animal knockout studies.

One of the surprising results was that VNB was more sensitive in this experiment than Affymetrix GeneChips, which showed no KHK expression in brain. This can be explained by high resolution of VNB on non-quantitative libraries, or, simply, by different methods of tissue preparation.

VNB Limitations and Scope

VNB is best suited to doing a first, very fast, pilot study prior to confirming the obtained expression data by experimental means. VNB accounts for splice variants only in the regions of alternative splicing. As other expression profiling methods, VNB results depend on the way a tissue was prepared (e.g. microdissected vs. the entire organ). Compare to microarrays VNB has a speed advantage only while using existing EST data. Although some groups do make new EST libraries specifically to derive expression profiles (Bodymap, http://bodymap.ims.u-tokyo.ac.jp/), microarrays allow for higher throughput in new experiments. Dynamic range of the method is dependent on the total number of available ESTs. While current number of ESTs in dbEST does not allow obtaining quantitative profiles for rare transcripts, VNB has a high qualitative resolution owing to normalized libraries. Finally, most ESTs in dbEST were obtained by oligo-dT priming in order to represent the cell's mRNA population. This method misses several recently discovered classes of regulatory RNA (Eddy**,** 2001) that do not have a poly-A tail (although the same is true for cDNA and currently commercially available oligo arrays).

Implementation

VNB is implemented as PERL script. Time required to obtain all accession numbers of the ESTs corresponding to the cDNA of interest for a high abundance gene (>1000 ESTs in dbEST) is 1-2 minutes on 1 GHz Intel Pentium III processor running LINUX. The time for obtaining expression profiles will be reduced farther after complete automation of the tool. Complete automation includes assembling gene-specific sets of ESTs for every known gene, automatic verification of ESTs that are duplicate reads, parsing for relevant tissue and library construction information, and reformatting dbEST into PostgreSQL object-relational database with web interface (public access is planned for the fall 2002). The database queries will follow basic biological rationales. The complete tool in its initial state will allow obtaining expression profiles for a known genes and novel sequences as well as tissue 'fingerprints'.

References

1. Adelman R.C., Ballard F.J., Weinhouse S. (1967) J. Biol. Chem. 242, 3360-3365.
2. Bais R., James H.M., Rofe A.M., Conyers R.A. (1985) Biochem. J. 230, 53-60.
3. Bergbauer K. et al. (1996) Dev. Neurosci. 18, 371-379.
4. Eddy S. (2001) Non-coding RNA genes and the modern RNA world. Nature Reviews Genet. 2, 919-929.
5. Erdos P., R'enyi A. (1970) On a new law of large numbers. Jour. Anal. Math. 23:103-111.
6. Funari V. (2001) Novel computational and classical molecular biology approaches to discovering alternative sites of fructose metabolism in mammals, Ph.D. thesis, unpublished.7. Funari V.A., Leyfer D. et al. (2000) Expression Profiling using the Expressed Sequence Tag (EST) Database for Comparative Physiology and Metabolism. In Recent Research Developments in Comparative Biochemistry & Physiology, (S.G.Pandalai, Ed.) Transworld Research Network, Kerala, India. 1, 13-30.
8. Haverty P. (Personal Communication), 2001.
9. Hayward B.E., Bonthron D.T. (1998) Eur. J. Biochem. 257, 85-91.
10. Leyfer D., Funari V. et al. A Novel Algorithm to Derive Unique Regions in Gene Sequences Is Utilized for in-silico EST Expression Profiling. Bioinformatics, manuscript in preparation.
11. Peri S., Ibarrola N. et al. (2001) Common pitfalls in bioinformatics-based analyses: look before you leap. Trends Genet, Issue 9. 1 September 2001. 17, 541-545.

Genome-wide expression profiling of *ESCHERICHIA COLI* W3110: MICROARRAY AND STATISTIC ANALYSIS OF HEAT SHOCK REGULONS

*[1,2] **Ozoline O.N.**, [2] **Fujita N.**, [2,3] **Ishihama A.**

[1] Institute of Cell Biophysics, RAS, Pushchino, 142290, Moscow Region, Russia
[2] National Institute of Genetics, Mishima, Shizuoka 411-8540, Japan
[3] Nippon Institute for Biological Science, Ome, Tokyo, 198-0024, Japan
e-mail: ozoline@icb.psn.ru

Key words: *E.coli, transcription, functional genomics, heat shock, microarrays*

Resume

Motivation: Availability of totally sequenced genomes allowed a possibility to use high-throughput approaches for annotation of the regulatory gene networks. *Escherichia coli* as a mostly studied organism provides the best opportunity to reveal new gene members of the hitherto identified regulons and to classify the genes with yet unknown functions.
Results: Microarray approach has been used to estimate the expression levels of about 4,000 genes of *E. coli* W3110 grown in minimal media and to characterize global changes in the transcriptome upon short-time exposure to high temperature. Two sets of the heat-shock responding genes, one enhanced and another repressed, have been identified and characterized in terms of the possible regulatory sequences.

Introduction

The total set of structural genes in *E. coli* MG1655 that was predicted from the complete genome sequence (Blattner et al., 1997) allowed a possibility to analyze their expression pattern at once using the microarray techniques. The gene expression data for several regulons and stimulons are presently available, including the comparative profiling of *E. coli* K12 MG1655 growing in LB or minimal media at the exponential, transition and stationary phases (Selinger et al., 2000, Wei et al., 2001a); after heat-shock treatment in rich media (Richmond et al., 1999); after exposure to hydrogen peroxide (Zheng et al., 2001); under the control of IHF (Arfin et al., 2000) and NtrC (Zimmer et al., 2000). Influence of SdiA amplification on the global gene expression was studied for *E. coli* RFM443 grown in LB media (Wei et al., 2001b). The complete genome sequence has also been determined for *E. coli* W3110 and expression profiling has been studied in response to physiological and genetic changes that affected tryptophan metabolism in both rich and minimal media (Khodursky et al., 2000). Extremely high sensitivity, *i.e.* up to 0,2 RNA molecules per cell (Selinger et al., 2000). of the microarray approach is, however, accompanied with high level fluctuation in the registered expression levels. To overcome this problem we tried to estimate typical deviations in the expression efficiencies of each E.coli gene at standard conditions. The data thus obtained were applied to characterize temperature-dependent alterations. Two sets of the genes differently responding to the temperature up-shift were further characterized in terms of the possible regulatory sequences located upstream from the corresponding open reading frames.

Methods and Algorithms

E. coli W3110 was grown in M9-glucose media. At the cell density of 60 Klett units, the culture (200ml) was divided into two parts. One half was transferred to the pre-warmed 47°C flask, whereas another half was continued to grow at 37°C. After 15 min incubation, cells were harvested, and total RNA was isolated using QiaGen RNeasy kit. Contaminating DNA was removed by DNase 1 treatment. Cy3 (control) and Cy5 (experimental) cDNA libraries were prepared using AMV-RT-XL (Takara), random primers and corresponding fluorescent analogs of UTP. Hybridization with DNA chips (Takara), which contain a total of 4028 DNA spots from 4390 open reading frames predicted for the entire genome, was performed for 16 h at 65°C. Microarrays were scanned with an Affimetrix laser scanner and the intensities of hybridized Cy3 and Cy5 were independently quantified by ImageQuant (Molecular Dynamics). Background correction was achieved by measuring the fluorescent intensity of the chip regions outside the DNA spots. Averaged signals detected for the spots containing unrelated DNA, *i.e.,* calf thymus DNA, human TFR or human beta actin, were used as negative controls. The data of 8 independent experiments were used to estimate standard deviation (std) in the expression efficiencies of particular genes.
Temperature dependent change in the relative level of particular mRNA was estimated on the basis of Cy5- to Cy3-intensity ratios within one and the same spot. Only those signals, which exceeded the background level for at least 3 std

* Corresponding author

were taken into account. Normalization was performed using an average Cy5/Cy3 value for the genes that were not affected by the temperature up-shift. One std of Cy3 or Cy5 intensities registered within the set of control spots was used as a minimum value to characterize the induction or the repression levels for genes, which are expressed only at 47°C or 37°C.

GeneSpring software (Silicon Genetics) was used to classify the up-regulated or down-regulated genes and to find potentially regulatory sequences in their upstream regions.

Results and Discussion

Despite the high sensitivity of the DNA chip method, which is obviously accompanied with fluctuation in the expression pattern, we found that the majority of individual RNA-products exhibited rather constant levels within the whole set of mRNAs. Standard deviations in the expression levels estimated for 3468 genes as a percentage to the corresponding mean values vary in the range from <1% to 119%, giving an average std for the whole set ~30%. The highest variability at standard conditions was registered for a group of genes, including *appA, appB, appC, BtuE, cheB, cheY, cheZ, flgC, flgD, flgE, flgF, flgL, flgM, fliF, fliM, fliS, fliZ, flxA, FruB, gadA, hdeA, hdeB, hdeD, hemN metA, metL, pheA, pheS, potF, pspB, slp, tap, tar, trpA, trpC, trpE, trpG, ybdL* and *yhiE*. The values of std for the percentage of detected RNAs of these genes were comparable with corresponding mean values (120-80%). Expression levels of slightly less than 300 genes vary within 80-50% from their average values.

Short-time heat shock treatment at 47 ^{O}C of the *E. coli* strain W3110 induced alteration in the expression profile of cellular RNAs. A total of 245 RNA species exhibited 2-50 fold increase while transcription of a total of 189 genes decreased. At least a part of the induced genes belongs to the sigma-H regulon. Thus, from 70 genes, heat-induced in a rich media (Richmond et al., 1999) and spotted on used DNA chips, only 4 (cycA, yjeH, yahB, hflX) were not heat-induced in our experiments. On the other hand, among the genes that were identified as up-regulated only in our experiments, the highest induction (47°/37° ratio > 5 and increase for more than 7,6 std) was observed for *baeR*, *chpS*, *cutA*, *cydB*, *feoA*, *relB*, *relE*, *uxaA, uxaC*, *uxuA*, *uxuB, uxuR*, *yaiB, yaiI, yaiY*, *ybdQ*, *ycjX*, *ydaL*, *ygdI*, *ygiW*, *yhfZ*, *yjbY*, *yrfE* and *zntR*. From 21 reliably down-regulated genes in a rich medium (Richmond et al., 1999), which were spotted on our microarrays, 5 (*endA*, *tsx*, *flgE*, *codA* and *pflB*) did not show the same changes. Among the genes that were down-regulated only in our conditions, the largest difference (37°/47° ratio > 5 and decrease for more than 4.6 std) has been detected for *appA*, *appC*, *aroA*, *aroF*, *grxB*, *hdeA, hdeB, hdeD*, *hisC, hisD*, *livJ*, *metE*, *ompF*, *pheT*, *shiA*, *rplB*, *rplC*, *rplD*, *rplI*, *rplV*, *rplW*, *rpsC*, *rpsF*, *rpsJ*, *rpsP*, *thiG*, *thiH*, *trpC*, *tyrA*, *tyrB*, *ybaS*, *ydiJ*, *yfiB* and *yojH*. The sets of responding genes thus identified may, therefore, contain some species, particularly required to maintain cell growth in minimal media.

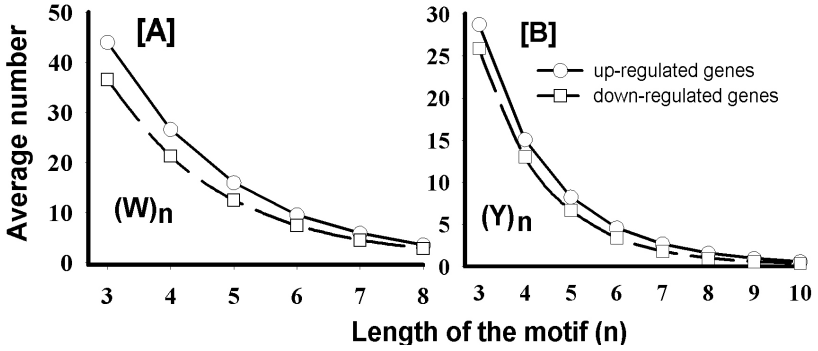


Fig. An average number of $(W)_n$, [A] and $(Y)_n$, [B], 20/250 base pair upstream from up-regulated (circles) or down-regulated (squares) at 47°C genes for the motifs of different lengths (n).

The set of up-regulated genes besides the previously annotated members of the heat- shock regulon contains the genes induced by hydrogen peroxide, the stationary phase-specific genes, the genes coding components of transcription and replication machineries, the isomerases and the proteases. Another set is mainly composed of translation associated genes, whose depression under stress conditions is expected. The expression is enhanced for at least 22 species of the transcription regulatory protein, including *baeR*, *cbl*, *copR*, *cytR*, *exuR*, *gcvA*, *gntR*, *greA*, *hepA*, *himD*, *malI*, *marA*, *mlc*, *mtlR*, *narP*, *ompR*, *phoB*, *phoP*, *rstA*, *sdiA*, *soxR*, and *uxuR*, while the expression was reduced only for 3 transcription factors, *evgA*, *hupB*, *nagC*. Thus, some of the heat-induced genes herewith identified may be attributable to the increased level of transcription factors, even though an average mRNA level does not directly correspond to the quantities of corresponding polypeptide.

After analysis of the upstream sequences of 20-250 base pairs in length for the genes that were affected by the temperature up-shift, we found that two groups of the genes, *i.e.* up-regulated and down-regulated, differ in some aspects. The up-regulated genes contain TAATT, AAATTT, CTTTT and some others as potentially regulatory sequences. The frequency in their presence upstream from down-regulated genes does not significantly differ from the whole genome. The down-regulated genes contain CGCAAAC, which is practically absent in the first set. All these motifs are identified as over-

presented when the whole set of DNA sequences upstream from open reading frames has been analyzed. More peculiar statistic analysis of the DNA sequences upstream from the differentially responding genes indicated that the frequency in the presence of $(W)_n$ (W=A=T) or $(Y)_n$ (Y=T=C) ($3 \leq n \leq 15$) is higher for the up-regulated genes (Fig.), providing a possibility that some transcription regulatory mechanisms might be mediated by the structural features of this type of DNA sequences. Alternatively these motifs may be targets for interaction with some specific or non-specific factors, which may be involved in stabilization of transcription initiation complexes at high temperatures. Thus, for instance, the RNA polymerase α subunit that efficiently interacts with the minor groove of A/T-rich DNA may selectively stabilize transcription complexes at the genes, which contain A/T-traces.

Acknowledgement

This work was supported by Ministry of Education, Science, Sports and Culture of Japan, the Core Research for Evolutional Science and Technology Corporation and the Russian Foundation for Basic Research (grants 00-04-48132, 01-04-97006)

References

1. Arfin S.M., Long A.D., Ito E.T., Tolleri L., Riehle M.M., Paegle E.S., Hatfield G.W. (2000) Global gene expression profiling in *Escherichia coli* K12. J. Biol. Chem. 275, 29672-29684.
2. Blattner F.R., Plunkett III,G., Bloch C.A., Perna N.T., Burland V., Riley M., Collado-Vides J., Glasner J.D., Rode C.K., Mayhew G.F., Gregor J., Davis N.W., Kirpatrick H.A., Goeden M.A., Rose D.J., Mau B., Shao Y. (1997) The complete genome sequence of *Escherichia coli* K-12. Science. 277, 1453-1462.
3. Khodursky A.B., Peter B.J., Cozzarelli N.R., Botstein D., Brown P.O., Yanofsky C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptofan metabolism in *Escherichia coli*. Proc. Natl Acad. Sci. USA. 97, 12170-12175.
4. Richmond C.S., Glasner J.D., Mau R., Jin H., Blattner F.R. (1999) Genome-wide expression profiling in *Escherichia coli* K12. Nucl. Acids Res. 27, 3821-3835.
5. Selinger D.W., Cheung K.J., Mei R., Johansson E.M., Richmond C.S., Blattner F.R., Lockhart D.J., Church G.M. (2000) RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. Nature biotech. 18, 1262-1268.
6. Wei Y., Lee J.-M., Richmond C., Blattner F.R., Rafalsky J.A., LaRossa R.A. (2001a) High-density microarray-mediated gene expression profiling of *Escherichia coli*. J. Bacteriol. 183, 545-556.
7. Wei Y., Lee J.-M., Smulski D.R., LaRossa R.A. (2001b) Global impact of SdiA amplification revealed by comprehensive gene expression profiling of *Escherichia coli*. J. Bacteriol. 183, 2265-2272.
8. Zheng M., Wang X., Templeton L.J., Smulski R.R., LaRossa R.A., Storz G. (2001) DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. J. Bacteriol. 183, 4562-4570.
9. Zimmer D.P., Soupene E., Lee H.L., Wendisch V.F., Khodursky A.B., Peter B.J., Bender R.A., Kustu S. (2000) Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. Proc. Natl Acad. Sci. USA. 97, 14674-14679.

MICROARRAY IMAGING DATA READER SPOTVIEW

**** Milanesi L., Rizzi R.***

Institute of Biomedical Technology CNR-ITB, Milan, Italy, e-mail: milanesi@itba.mi.cnr.it
*Corresponding author

Key words: *microarray, mathematical model, gene, computer analysis*

Resume

Motivation: The novel technology of microarray expression assay investigates the expression behavior of a high number of genes at the same time. Target sequences are arranged onto small membranes according to schemes more or less regular. The target sequences are then driven to react with the gene product extracted from the cellular populations to be studied, in order to produce signal spots whose intensity is directly proportional to the expression level of the corresponding sequence in the target panel. The produced signal is then digitized in order to obtain digital images to be processed and analyzed.
It is therefore necessary to localize the signal spot on the image for reading the contained expression data. By defining areas of interest for the expression spots and for the surrounding backgrounds, an integration value and a value of local noise for each signal band are obtained. The different acquisition steps lead furthermore the resulting expression images to contain several artifacts and errors (background noise, overlapping, signal saturation, etc.) to be taken into account if reliable and correct expression data must be reached.
Results: A program, called SpotView, has been developed in order to evaluate these aspects. Regular grids of interest areas are placed on the image, following a semi-automatic procedure of position adjustment, and the signal for each band is retrieved. The grid positioning, the correct spot value and the background subtraction are very important for the preliminary pre-processing of expression images. An extensible simulator has been also developed to obtain the basis for producing synthetic image examples on which to accomplish studies.
Availability: contact milanesi@itba.mi.cnr.it

Introduction

The study of cellular gene expression profiles is becoming more and more important within the molecular biology, the biochemistry and the medicine, as it allows shedding light on a series of fundamental factors regulating the life of every organism, from simple organisms like yeast to human being. In the last years a new methodology, based on microarray, has been developed for investigating the cell behavior in comparison to the expression of a high number of genes. The main goal of this methodology is understanding the role of genes by processing simple digital images. One of the advantages of this technique is that of being highly parallel with the possibility of making it automatic. A microarray expression assay, like traditional genetic methods, utilizes the hybridization reaction between complementary nucleotide chains. Nucleotide sequences are deposited, according to a predefined scheme, onto small membranes (i. e., nylon and glass) in order to obtain a sort of genetic microchip (also called genetic microarray). Such sequences are extracted from genomic libraries (i.e., cDNAs or ESTs (Chen et al., 1998)) or synthesized, in accordance with known and pre-arranged sequences, by means of photolithography techniques: synthetic oligonucleotides (Wodicka et al., 1997). A known amount of mRNA is extracted from the cells to be studied, it is labeled by color-forming substances and is submitted to hybridization reaction with the panel of genes of the microarray. When hybridization occurs, chemical reagents lead the color-forming substances "to develop" the image signal. After washing the membrane, the signal is read and digitized through scanning devices. Signal digitization can occur through flatbed scanners, drum scanner, color video cameras or digital cameras with stereomicroscope (Chen et al., 1998), confocal scanners (Wodicka et al., 1997; Lockhart et al., 1996) or finally through laser-inducted fluorescence scanning devices (Schena et al., 1996). The standard microarray format is 16-bit TIFF. Signal intensity in a particular target position is then expected to be directly proportional to the amount of the cellular mRNA which reacted with the deposited sequence and therefore proportional to the expression level of the deposited gene sequence in the cellular population.
Processing of resulting expression patterns turns out very crucial, on one hand, to associate to each target gene an expression value within the studied cellular population and, on the other hand, to obtain the expression profile of that gene in different experiments. Normalization and quantification of expression data is fundamental to produce expression values and profiles that are reliable in comparison to well-known internal and external controls.

Overview of microarray image problems

Processing of expression patterns is the main task towards the determination of the cellular expression profile; therefore it's necessary to take into account which factors contribute to the signal production, from the experimental protocol to the signal digitizing; several systematic and accidental errors make results noisy and the experiment will not be reproducible by different operators.

Signal intensity is directly proportional to the mRNA amount and to the target amount and the risk of signal saturation is present; (Chen et al., 1998) reported the results for two internal control genes with expression ratio tending to the actual ratio in the presence of low mRNA amounts even though low expression targets often mix with the background noise.
The great disturbing element in a microarray expression image is the noise, that is composed mainly of noise by residual hybridization, background noise due to the substrate material and background noise due to the scanning process and to the scanning system optics. The residual hybridization noise is due to the fact that a minimum hybridization reaction, between the target and the studied sample, always exists. The noise due to the substrate material is more properly a background noise and it is present in the "phosphoring imaging" assays. Presence of organic material, such as dust, skin, finger oil, or sample preparation substances may produce a fluorescent signal that is retrieved if its wavelength is in the range utilized by the scanning instrument. Noise deriving from the scanning process is photon statistical noise, electronic noise, fluorescence of the optical components of the scanner, excitation reflected and scattered by the membrane substrate.
Overlapping among contiguous spots limits the microarray panel density, a high target density would impede the user to distinguish signal areas of different targets.
All the errors affecting microarray images reduce the image quality from the point of view of the data retrieval and the reliability of the final expression results. Every problem and error, implicit in the microarray assay process, must be accurately taken into account in order to accomplish a correct reading of the expression patterns and reliable data normalization.

Data representation for microarray reading

A program has been developed to this aim. Regular grids of interest areas are defined for the expression image and the band signal is obtained for each grid element. The user can define a preliminary positioning of the grid on the microarray image. At this point an automatic adjustment of the preliminary positioning is obtained by applying an iterative affine transformation to the grid until the optimal total integration is achieved. This procedure may encounter several problems depending on the various aspects of the image formation (from the experimentation protocol to the digitization of the signal). The local background is computed by multiplying the mean intensity value in the interest area for the background, by the interest area for the expression spots. In order to avoid neighboring light spots affect the local background value, the local background is substituted by the median of the neighborhood. Some statistics can be produced for attaching to each single spot a quality measure, i. e. standard deviation of the intensity, saturation percentage and percentage of pixels with intensity greater than or equal to one standard deviation above the median intensity in the interest area of expression.
The different acquisition steps lead the resulting images to contain several artifacts and errors (background noise, overlapping, signal saturation, etc.). Different reading formats and different types of filter can be used by SpotView. The resulting data from the final image analysis and the characteristics for the grid positioning can be saved in file formats ready to be imported in a relational database.
The first step for obtaining data of gene expression level, is reading raw image data. To accomplish this task, regular grids of interest areas are placed on the image and the signal for each band is retrieved. If the grid positioning follows the signal pattern of the image, the value of signal integration in the interest areas of the grid gives a raw measure of the expression level for the cellular populations bound to each band. This procedure is indicated for those expression images containing regular expression patterns and therefore derived from microarray constructed with arraying machines arranged as matrixes of pins. A reading grid is composed of interest areas laid out in columns and rows and its regularity simplifies its definition and positioning. More independent reading grids can be placed on a image. Each element of a reading grid is composed of one signal interest area and one area for estimating the background noise. Geometric parameters define the grid through number of elements per row, number of elements per column, spacing along a row, spacing along a column, shape and dimension of the signal interest area and shape and dimension of the noise estimate area.
An integration data I_b, for the targets in the panel, will be produced for each band of signal b by integrating the signal in the grid interest areas As:

$$I_b = \sum_{i \in As} s(i,b) \tag{1}$$

The background noise n_b is expected to have a constant value in the signal interest areas. This value is estimated for each band b as the average of the signal intensity in the noise estimate area An of the grid, which contains C_n pixels:

$$n_b = \frac{1}{C_n} \sum_{i \in An} s(i,b) \tag{2}$$

The background noise N_b of a generic spot for the band b will be given by multiplying the mean noise n_b by the signal interest area As:

$$N_b = n_b \times As \tag{3}$$

By subtracting the estimated background noise N_b from the value obtained with (1), the integration data I_b is purified from noise to obtain the raw expression data E_b:

$$E_b = I_b - N_b \tag{4}$$

The noise estimate area should be circumscribed to the spot in order to avoid the signal of the neighboring spots being included in the estimate procedure. Light spots in the neighborhood affect and distort the local background value. Disturbs of neighboring light spots can be avoided by using the median of the neighborhood's background values.
Information about target deposition areas helps define the size of the signal interest area of reading grids. Before reading raw expression data the grid must be centered on the image pattern.
An a-priori knowledge of the image acquisition process help define and place the reading grid on the image. A semi-automatic procedure of grid positioning is required. To a preliminary manual approximate positioning follows an automatic position refinement based on the signal pattern, that leads the reading grid to its right place. The grid parameters are progressively updated until a sort of permanent equilibrium is reached. This last condition is reached by maximize (or minimize) a predefined grid-related measurement. The refinement procedure and therefore the final position are clearly affected by the chosen group of movements intended as number, type and order, by the selected signal band and finally by the predefined measurement leading the entire process. A refinement procedure can also be applied to each single part of the grid.
At this point, the reading grid is centered on the signal pattern and raw expression data can be retrieved.
Language and computer used. The following computer programs have been entirely developed in JAVA code for platform portability.

Acknowledgements

The work was supported by the EC IST 2001 32688 – ORIEL, CNR Functional Genomics and Oncology Over Internet O2I projects.

References

1. Chen J.J.W., Wu R., Yang P-C., Huang J-Y., Sher Y-P., Han M-H., Kao W-C., Lee P-J., Chiu T. F., Chang F., Chu Y-W., Wu C-W., Peck K. (1998) Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics. 51: 313-324.
2. Lockhart D.J., Dong H., Byrne M.C., Follettie M.T., Gallo M.V., Chee M.S., Mittmann M., Wang C., Kobayashi M., Horton H., Brown E.L. (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nature Biotechnology. 14: 1675-1680.
3. Schena M., Shalon D., Heller R., Chai A., Brown P.O., Davis R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. of the Natl Acad. of Sci. USA. 93: 10614-10619.
4. Wodicka L., Dong H., Mittmann M., Ho M-H., Lockhart D.J. (1997) Genome-wide expression monitoring in Saccharomyces cerevisia. Nature Biotechnology. 15: 1359-1367.

PROBLEMS OF CONTROL OF GENE NETWORKS IN A SPACE OF STABLE STATES

[1] Latypov A.F., [1] Nikulichev Yu.V, [2] Likhoshvai V.A., [2] Ratushny A.V., [2] Matushkin Yu.G., [2] Kolchanov N.A.

[1] Institute of Theoretical and Applied Mechanics, SB RAS, Novosibirsk, Russia, e-mail: latypov@itam.nsc.ru
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
*Corresponding author

***Key words**: gene network, stable state, nonlinear and linear programming, sensitivity matrix*

Resume

Motivation: Mutations occurring in genes result frequently in impaired operation of gene networks, causing a diversity of pathologies. Of highest importance is to find the optimal factors and strategies for correcting their functions. This necessitates development of the methods allowing the gene network operation to be controlled.
Results: This work describes statements of problems of control in a space of stable states with reference to gene networks reduced to problems of linear and nonlinear programming. A method for solving nonlinear problems is proposed. A sensitivity matrix for the simplest model simulating cholesterol biosynthesis is constructed and analyzed.

Introduction

Gene network is a set of concertedly expressed genes controlling performance of a particular body function. A group of specific genes together with the following elements form the core of a gene network: (1) proteins encoded by the genes in question; (2) pathways of signal transduction from cell membranes to cell nuclei, providing activation or inhibition of gene transcription; (3) negative and positive feedbacks, either stabilizing gene network parameters at a certain level or, on the contrary, deviating them from the initial value, thereby switching the system into a new functional state; and (4) low-molecular-weight components, switching the gene network function in response to external stimuli (hormones and other signal molecules), energy carriers, various metabolites, etc. (Kolchanov et al., 2000).
Normal gene network operation requires a concerted interaction of all its components in both space and time domains as well as its ability to adequately receive and process the external signals. Impairment of any component may lead to a certain degree of impairment in the gene network operation, resulting in various pathologies. This arises the problem of adequate correction of the damaged gene network to restore its normal (or close to normal) operation. However, gene networks may comprise hundreds and thousands of components interconnected in an intricate nonlinear manner. Therefore, the correcting effects, selected from a list of allowed corrections, may cause certain unfavorable side effects. Thus, any particular correction of the gene network operation may be considered permissible only if it is capable of restoring the major gene network function with minimal side effects. Summing up, the problem of searching for optimal intervention into gene network operation is a problem of control in a broad sense.
An important specific feature of the problem of gene network control is the necessity of solving it in two stages, that is, it is first necessary to find out whether a state required is existing. This brings about the problem of control in the space of stable states (PCSS). Then, if the answer is positive, the problem of searching for optimal realization of the state required in the dynamic process is solved.
A schematic representation of correcting the gene network function on the surface of stable states is shown in Fig. 1. The gene network in a state of pathology A is subjected to controlling effects and transferred to the normal state B. Not all the effects are allowed (crossed out here), as they lead to adverse side effects.
The problems of control have not been yet stated in the context of gene networks. The goal of this work is to fill partially this gap. Here, we are formulating two problems of control of gene network stable states. We are also considering certain approaches to their solution based on considering the sensitivity matrix and solving problems of control “locally”.

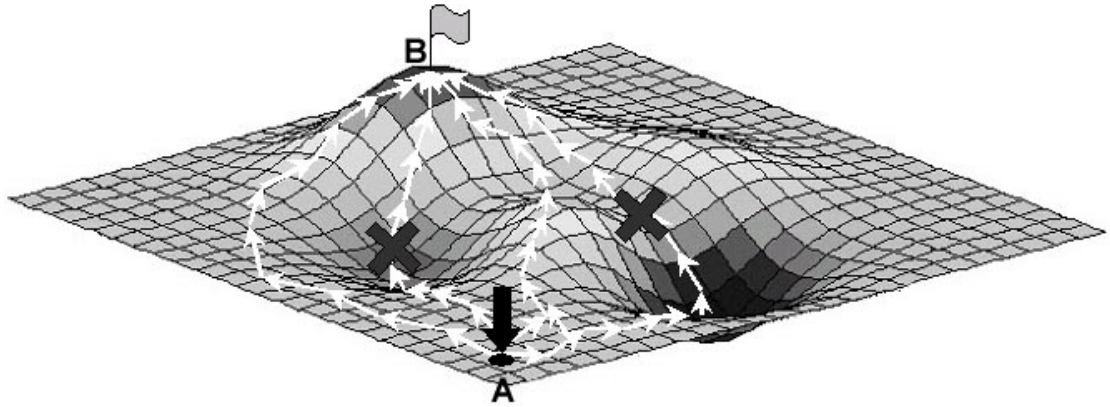


Fig. 1. Qualitative illustration of a transition of the system from pathological state A to the normal state B. Note that tot all the trajectories from A to B are permissible.

Problems of control of stable states

In the context of modern concepts, the operation of a gene network (GN) is described with a system of ordinary differential equations (ODE; Likhoshvai et al., 2001):

$\dot{V} = f(V, a)$,

where **V** is the vector of parameters of a GN state and $\boldsymbol{a}$, the vector of internal parameters. Let the values $\mathbf{a} = \mathbf{a}_*$ correspond to the normal state of gene network. Let GN in the normal state maintain a certain equilibrium state. Formally, this means that the system $f(V,a) = 0$ allows at least one nontrivial solution to exist in the range of positive values V: $V = V_*$. Let us consider the state V_* as normal (basic). Let us designate $w = \frac{V}{V_*}$, $\alpha = \frac{a}{a_*}$, $\tau = \frac{\mathrm{t}}{\mathrm{T}}$, where T is the characteristic time of the process. Then, (1) may be expressed in a dimensionless form as

$$\varepsilon \frac{\mathrm{d}w}{\mathrm{d}\tau} = \mathrm{T}\varphi(w, \alpha).$$

Let the totality of parameters $\boldsymbol{\alpha}$ be divided into three following groups (this partition depends on the particular situation studied):

1) α_k, $k \in K$ is the totality of assigned numbers of the parameters displaying changed values that arouse due to mutations;

2) α_l, $l \in L$ is the totality of assigned numbers of the parameters used as controls; and

3) α_m, $m \in M$ are the rest parameters that are remaining unchanged, that is, $\alpha_m = 1$.

Let us distinguish two statements of the problem of control of stable states.

If the procedure used for selecting the starting point of the control (the stable point whereat we are intervening the GN operation) and the trajectory for reaching the basic point within the region specified are not essential, PCSS is formulated as follows:

Problem No. 1. Let us specify

1) Certain values $\alpha_k = \overline{\alpha}_k$ of the parameters from group ***K;***

2) The region of admitted values $\alpha_l \in D_l = \left\{\alpha_l : \alpha_l^{\min} \le \alpha_l \le \alpha_l^{\max}\right\}$ of the parameters from group ***L;***

3) Neighborhood of the basic point $\mathrm{B}_\Delta = \left\{w : 1 - \Delta_{\min} \le w \le 1 + \Delta_{\max}\right\}$ and the functional $F = \frac{1}{2}\sum_l \beta_l (\alpha_l - 1)^2$, β_l are weight coefficients $\sum_l \beta_l = 1$.

It is necessary to determine such $\widetilde{\alpha}_l$ that $\varphi(\widetilde{w}, \overline{\alpha}_k, \widetilde{\alpha}_l)\big|_{\alpha_m = 1} = 0$ (stationary condition), $\widetilde{w} \in \mathrm{B}_\Delta$ (condition of occurrence in the region), and $F(\widetilde{\alpha}_l) = \min\limits_{\alpha_l \in D_l} F(\alpha_l)$.

The functional ***F*** characterizes minimal deviations of the parameters from their basic values and, therefore, it is likely that minimal "expenditure" effects will be required for achieving α_l.

If PCSS solution requires taking into account the selection of starting point and the trajectory for reaching the basic point within the region specified, we are coming to the formulation of problem No. 2.

Problem No. 2. The stationary point $\overline{w}$ from equations $\varphi(\overline{w},\overline{\alpha}_k)\big|_{\alpha_l=1,\ \alpha_m=1}=0$ is determined in addition to the statement of problem No. 1. It is necessary to determine the sequence α_l^i, $i=0,1,\ldots$, such that $\varphi(w^i,\overline{\alpha}_k,\alpha_l^i)\big|_{\alpha_m=1}=0;$ $\phi(w^{(i)})\le 0;$ $\lim\alpha_l^{(i)}=\tilde{\alpha}_l,\ \alpha_l^{(0)}=1;\ \lim w^{(i)}=\tilde{w},\ w^{(0)}=\overline{w};$ and $\tilde{w}\in \mathrm{B}_\Delta;\ F(\tilde{\alpha}_l)=\min\limits_{\alpha_l\in D_l}F(\alpha_l)$. Here, the function $\phi(w)$ specify the region of admitted trajectories in the space of states.

The software package Poisk (Latypov, Nikulichev, 1985) may be used for solving problems № 1 and 2. The package has been essentially revised, adapted to the problems of gene network control, and realized in the Delfi–4 environment.

Solving of PCSS requires frequently obtaining of the following information:

- Determining the necessary accuracy for specifying the vector of internal parameters **a;**
- Determining the rational composition of control parameters from the totality of **a;**
- "Calibrating" the parameters according to their significance; and
- Evaluating the adequacy of the GN model considered.

The most optimal method to solve such problems is application of **sensitivity matrix.** Sensitivity matrix may be constructed at any stationary point. Let us describe the procedure of its construction.

For definiteness, let us consider the basic point $\alpha=1,\ w=1$. The initial equations at the stationary point take the following generalized form

$\varphi(w,\alpha)=0,\ \varphi=\{\varphi_i, i=\overline{1,n}\},\ w=\{w_i,\ i=\overline{1,n}\},\ \alpha=\{\alpha_j,\ j=\overline{1,m}\}$,

and variation form $J\cdot \mathrm{x}=\mathrm{b}=-C\cdot \mathrm{p},\ \mathrm{x}=\delta w,\ \mathrm{p}=\delta\alpha,\ J=\dfrac{\partial\varphi}{\partial w},\ C=\dfrac{\partial\varphi}{\partial\alpha}.$

Then, the solution is as follows: $x=D\cdot p,\ D=J^{-1}\cdot C=\{\gamma_{ij}\},\gamma_{ij}=\dfrac{x_i}{p_j}$

For a simplified variant of the mathematical model simulating cholesterol biosynthesis regulation (Ratushny et al., 2000), containing 10 equations and 38 parameters, we constructed the matrix of coefficients relating the variations in internal parameters of the mathematical model to the changes in stationary state parameters (sensitivity matrix). Analysis of this matrix allowed us to distinguish three following groups of the parameters used in the model in question: (1) the group of a high effect (comprising only one parameter) with a coefficient of influence amounting to $\sim 10^2$; (2) the group of a moderate effect (six parameters) with a coefficient of influence of ~ 1; and (3) the group of a weak effect (all the rest parameters) with a coefficient of influence equaling $\sim\le 0.5$. A number of parameters of the third group, falling into a separate subgroup, display a very weak effect on the parameters determining the state of the system. In addition, certain parameters exhibited moderate effects only on individual elements of the system without any noticeable effect on the rest.

The analysis performed suggests that depending on the problem solved, the sensitivity matrix alone allows the preliminary conclusions on certain possibilities in controlling the system to be made. Moreover, the sensitivity matrix gives useful information while solving the problem of determining the values of model's variables from measurements of the stationary states of the system. For example, to increase the accuracy of estimating a particular variable, it is appropriate to measure the elements, the coefficients of influence on which with reference to the variable in question is ~1, as the error increases multifold at a low coefficient due to the inversely proportional dependence, whereas time resolution might be inaccessible at a high value of the coefficient.

With reference to PCSS, the sensitivity matrix should be considered when the optimal control is searched for within a small neighborhood of the basic (normal) state (local control). In this case, the PCSS is reduced to solving a sequential set of liner programming problems. A linear programming problem is solved in two following stages. At the first stage, a point x^* satisfying the condition for occurrence in the region specified is determined:

Problem L1.

$$x_i=\sum_k\gamma_{ik}p_k+\sum_l\gamma_{il}p_l;\ |x_i|\le\Delta_i;$$

$$F=\sum_i c_i x_i\Rightarrow\min_{p_l\in P_l};\quad c_i\in\{-1,0,+1\}$$

where x_i is changes in the parameters of the state caused by mutation (the sum over k) and by control (the sum over l) and Δ_i specifies the admissible region. The functional F provides that the point occurs in the region specified. Let us designate the solution of Problem L1 as $p_l=p_l^*$. Then, the second stage problem is solved.

Problem L2.

$$x_i = \sum_k \gamma_{ik} p_k + \sum_l \gamma_{il} p_l;\ |x_i| \le \Delta_i;$$

$$F^* = \sum_l Sign(p_l^*)\,\beta_l p_l \Rightarrow \min_{p_l \in P_l}$$

Conclusion

Described here are the statements of the problems of control in a space of stable states applied to gene networks that are reduced to problems of linear and nonlinear programming and the method for solving nonlinear problems. The sensitivity matrix for a simplest model of cholesterol biosynthesis is constructed and analyzed.

Acknowledgements

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

References

1. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000). Gene Networks. Mol. Biol. (Mosk.). 34:449-460.
2. Latypov A.F., Nikulichev Y.V. (1985), Specialized group of optimization programs. Preprint ITAM SB AS USSR No. 15-85, Novosibirsk.
3. Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. (2001). A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk.). 35(6):1072-1079.
4. Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. (2000). Mathematical model of cholesterol biosynthesis regulation in the cell. Proc. Second Intern. Conf. on Bioinformatics of Genome Regulation and Structure, Novosibirsk, 1:199-202.

A METHOD OF SOLVING PROBLEMS OF OPTIMAL CONTROL IN DYNAMICS OF GENE NETWORKS

[1] Latypov A.F., [1] Nikulichev Yu.V., [2] Likhoshvai V.A., [2] Ratushnyi A.V., [2] Matushkin Yu.G., [2] Kolchanov N.A.

[1] Institute of Theoretical and Applied Mechanics, SB RAS, Novosibirsk, Russia, e-mail: latypov@itam.nsc.ru
[2] Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia
*Corresponding author

***Key words**: gene network, mathematical model, dynamics, optimal control*

Resume

Motivation: Dysfunction of gene networks caused by mutations in genes is often the reason for development of pathologies in an organism. Optimal actions on a gene network can recover its normal functioning. To seek for such actions, one has to develop methods of control of the functioning dynamics of gene networks.
Results: Problems of optimal control of the functioning dynamics of gene networks are formulated in the paper.

Introduction

Generally, the functioning of a gene network is ensured by correlated interaction of all its chains both in space and time and by the capability of the gene network to adequately accept and treat external signals (Kolchanov et al., 2000). Dysfunction in an arbitrary chain can disturb the function of the gene network. In turn, this can be the reason for pathologies. In this case, there arises the problem of correction of gene network functioning in order to recover its normal (or close to normal) operation. However, gene networks consist of hundreds and thousands elements with complex nonlinear links. Therefore, actions chosen among the list of allowed actions may have some adverse side effects. Hence, a particular correction of dysfunction of the gene network can be considered as admissible if the use of this correction not only recovers the main function of the gene network but also minimizes side effects. Thus, the problem of searching for optimal interference into gene network operation is the problem of dynamic control in a wide sense.
Figure 1 shows a schematic of correction of gene network operation. From state A (pathology), the gene network is transformed to the normal state B by means of control action. Not all actions are allowed: some potential trajectories are crossed.

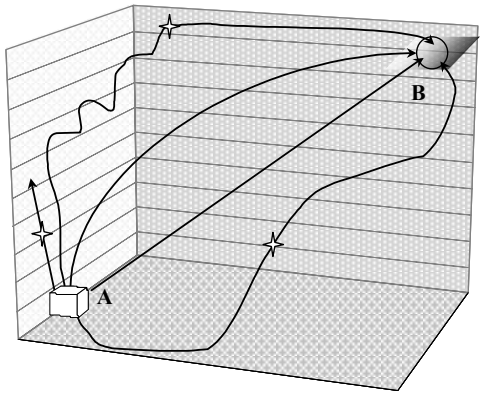


Fig. 1. Qualitive illustrtion of converting a system from a pathological state A to the normal state B. Not all trajectories are allowed.

An approximate solution of the problem of optimal control in a wide sense is based on the correspondence to the control of a finite-dimension search space represented by a given group of operations on elementary functions with a finite set of parameters. An adequate representation of the control should satisfy certain requirements; the most general requirements are the absence of induced oscillations, monotonicity, and necessary accuracy. Particular problems often impose additional requirements on the control, such as discontinuity, piecewise continuity, linearity, etc. An important factor is the simplicity: the minimum number of parameters sufficient for representation of control with prescribed properties, which is often responsible for computation time and resources needed. Multichain representations that started from cubic splines (Ahlberg et al., 1967) are most popular. Rational and exponential splines (Kvasov, Yatsenko, 1988) and methods of Bezier (Bezier, 1971) and Ferguson (Ferguson, 1964), which are rather popular and convenient for representation of physical regions and objects, do not always guarantee necessary isogeometric properties and, being not oriented to solving control problems, do not ensure sufficient simplicity of computations and minimum consumption of resources. The most convenient representation of the control, which possesses the above-described properties, is the *LL*-approximation (Aulchenko et al., 1998).

Dynamic control problems have not been formulated as applied to gene networks. The present work has to partly fill this gap. Problems of optimal control of the operation dynamics of gene networks are formulated here. Some approaches to their solution are also considered.

Problems of optimal control in gene network dynamics

Models used to study gene networks are characterized by the presence of piecewise-linear and discontinuous control functions. Parametric *LL*-approximation allows one to represent discontinuous functions, determining points of discontinuity with an arbitrarily high accuracy.

Within the assumptions used (Likhoshvai et al., 2001), functioning of a certain gene network (GN) is described by a system of ordinary differential equations (ODE)

$$\dot{\mathrm{V}} = \mathrm{f}(\mathrm{V}, \mathrm{a}). \tag{1}$$

Here **V** is the vector of the GN state and **a** is the vector of internal parameters determining the evolution of the GN state (interactions of elements of the system with each other and with the ambient medium, decomposition, generation, introduction from outside, ejection, etc.).

The steady state of the system satisfies the equations

$$\mathrm{f}(\mathrm{V}, \mathrm{a}) = 0. \tag{2}$$

Let $\mathbf{a}=\mathbf{a}_*$ correspond to $\mathbf{V}=\mathbf{V}_*$. This state is called normal (basic).

We denote

$$\mathrm{w} = \frac{\mathrm{V}}{\mathrm{V}_*}, \quad \alpha = \frac{\mathrm{a}}{\mathrm{a}_*}, \quad \tau = \frac{\mathrm{t}}{\mathrm{T}}. \tag{3}$$

(*T* is a characteristic time). Then (1) and (2) can be written as

$$\varepsilon \frac{\mathrm{d}\,\mathrm{w}}{\mathrm{d}\,\tau} = \mathrm{T}\,\varphi(\mathrm{w}, \alpha), \tag{1'}$$

$$\varphi(\mathrm{w}, \alpha) = 0. \tag{2'}$$

The presence of the parameters ε in (1') follows from the accepted condition that the norm of the Jacobi matrix at the initial point is of the order of unity.

The right-hand part of Equation (1) contains a term of the form $-w_i/T_i$. Therefore, to trace the fastest process with a required accuracy, one has to assume that $\max_i(\frac{\mathrm{T}}{\mathrm{T_i}}) = 1 \Rightarrow T = \min_i T_i$.

Let the set of parameters α be divided into three groups (this division depends on a particular situation under study):

1) α_k, $k \in K$ is the set of the numbers of parameters that define the disturbed state (mutation);

2) α_l, $l = L$, the set of the numbers of parameters used as control parameters; and

3) α_m, the remaining unchanged parameters, i.e., $\alpha_m = 1$.

Obviously, by virtue of normalization (3), Eqs. $\varphi(\mathrm{w},1) = 0$ have the solution **w**=1 (reference point).

Let the gene network dynamics be described by the system

$$\varepsilon \frac{dw}{d\tau} = \varphi(w, \lambda, U);\ w = \frac{V}{V_0};\ \lambda = \alpha_k \bigcup_{\substack{k\in K \\ m\in M}} \alpha_m;\ U \in D = \{U : U^{\min} \le U(\tau) \le U^{\max}\} \tag{4}$$

The parameters α_l are denoted here via the U – control (vector of dimension *P*). The control problem is formulated as follows.

Problem

The prescribed parameters are α_k determining the disturbed state, $\alpha_m = 1$, and *B*, which is a Δ_i -vicinity of the reference point.

The task is to transform the point from one state to another in the space of states by means of control by parameters of the group L under functional restrictions and the condition of functional minimization.

The problem is formally written as follows:

$$w(\tau_o) \Rightarrow^{u(\tau)} \Rightarrow w(\tau_k) \in B = \{w_i : |w_i - 1| \le \Delta_i, i = \overline{1,n}\};$$

$$\psi(w, \lambda, u) \le 0;$$

$$\varphi(\tau_k) = \varphi(w(\tau_k), \lambda, u(\tau_k)) = 0;$$

$$\lambda = \alpha_k \bigcup_{\substack{k \in K \\ m \in M}} \alpha_m; \tag{5}$$

$$u(\tau) \in D = \{u : u^{\min} \le u(\tau) \le u^{\max}\};$$

$$F(w_k, \tau_k) \Rightarrow \min_{u(\tau) \in D, \tau_k}$$

The condition $\varphi(\tau_k) = 0$ means retention of the point $w(\tau_k)$ in B for an infinitely long time. The functional F is assumed to be $F = \tau_k$ at the first stage.

Three basic variants of the problem may be considered.

The point $w(\tau_0)$ is a disturbed steady state.

The point $w(\tau_0)$ is the reference steady state.

The point $w(\tau_0)$ is an arbitrary nonequilibrium point.

In the second case, the motion is due to setting $\alpha_k \ne 1$.

The problem formulated belongs to the class of problems of nonlinear programming: minimization of a given functional under given nonlinear restrictions written as a system of equalities and inequalities. The efficiency of search algorithms in hyperspaces depends significantly on the manner the restrictions are taken into account. We use the method of constructing a single composite functional that controls deviations of restrictions with a prescribed accuracy. The composite functional is formed as follows:

$$\Phi = F\left[1 + \sum_{j=1}^{p} \delta_j \left(\frac{\phi_j}{\varepsilon_j}\right)^2\right] + \sum_{j=1}^{q} \delta_{j+p} \left(\frac{\varphi_j}{\varepsilon_{j+p}}\right)^2. \tag{6}$$

Here q and p are the numbers of functional restrictions in the form of equalities and inequalities, respectively, ε is an array whose components contain the required accuracies of satisfaction of restrictions of the corresponding equalities and inequalities, and δ is an array whose components are the penalty coefficients whose values are adjusted in the course of the search. It was proved (Latypov, 1974) that the estimate is $\delta \approx \varepsilon$ from the condition of a local minimum of Φ in R_n.

To solve the problem, one can use the software system Poisk (Search) (Latypov, Nikulichev, 1985). The software system is adapted to problems of gene network control and implemented in the Delfi-4 environment.

We describe the algorithm of solving the problem of nonlinear programming, which forms the basis of the Poisk system. The main blocks are as follows:

LUCH, an one-dimensional search combining the method of parabolic approximation and the method of golden section;

EPCP, a coordinate descent with a random choice of the number of the varied coordinate or the numbers of coordinate groups. In this case, the number of the coordinate with the least influence for the slope method is determined (Latypov, 1974) and a "rebound" (a small step behind from the local extremum found) is performed, which ensures stability of the search;

PCEVDG, a random search in a half-hypercone with a given direction of the axis (unit vector dfm) and half-angle θ; and

SHAR, a random search in a hypersphere with a given radius and center position (matrix descent, since a certain group of variables is "frozen").

The Poisk program is organized as a certain sequence of initiation of these blocks with the following adaptation of program parameters in the course of operation.

In the beginning of program operation, the search is performed with a low accuracy of satisfaction of restrictions and determination of the uncertainty interval in one-dimensional descents. The accuracy is later increased to prescribed values by an iterative method. Schematically, the operation algorithm of the Poisk program can be described as follows (the term "success" indicates that the block has operated rather efficiently).

1. Setting the initial values of program parameters, read-out of problem data, other preparatory operations, and calculation of the initial value of the functional.
2. EPCP with "rebound" and calculation of the vector dfm. If "success", go to Sec. 3, otherwise, to Sec. 4.
3. PCEVDG and go to Sec. 2.
4. SHAR. If "success", then calculate the vector dfm and go to Sec. 3, otherwise, go to Sec. 5.
5. If the "working" accuracies have not reached given values, perform the next iteration on refining the changes in the "working" parameters of accuracy and penalty coefficients and go to Sec. 2, otherwise, END.

As an example, we solve the control problem with the use of *LL*-approximation and Poisk program for an ODE system with one control function $u(t)$. For the system

$$\dot{y}_1 = y_2,\ \dot{y}_2 = \frac{u(t) - by_2^2}{y_3} - 1,\ \dot{y}_3 = -u(t) \qquad (7)$$

with the initial conditions $t = 0$: $y_1 = 0,\ y_2 = 0,\ y_3 = 1$ and a restricting condition of the form $y_3(t_k) - \mu = 0$, we have to determine the minimum of the functional $F = -y_1(t_k)$.

The varied parameter here is t_k, the value of the argument up to which integration of system (7) is performed. The exact solution of this problem has the following form (two points of discontinuity of the control):

$$u(t) = \begin{cases} u_{\max}, & y_3 > by_2^2 \cdot (1 + y_2), \\ \dfrac{by_2(2+3y_2)\cdot(by_2^2 + y_3)}{by_2(2+3y_2) + y_3}, & y_3 = by_2^2 \cdot (1 + y_2), \\ 0, & y_3 \le \mu. \end{cases}$$

(Bryson, Ho Yu-Shi, 1968). The solution obtained by the algorithm described is plotted in Fig. 2 for b=10.24 and μ=0.2. The difference of the solution obtained from that calculated by analytical formulas is less than 10^{-6}.

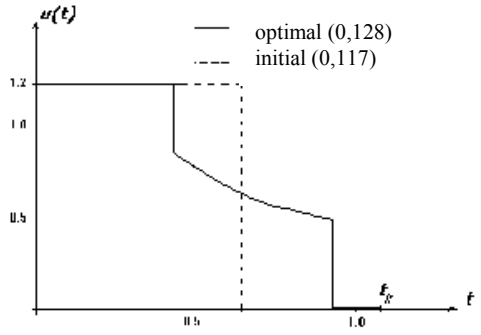


Fig. 2. An example of solving the control problem.

Acknowledgements

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

References

1. Ahlberg J., Nilson E., Walsh J. (1968). The Theory of Splines and Their Applications. London, Academic Press.
2. Aulchenko S.M., Latypov A.F., Nikulichev Yu.V. (1998). Construction of curves with the use of parametric polynomials. Zh. Vychisl. Mat. Mat. Fiz. 38:1967-1972.
3. Bezier P. (1971). Example of an existing system in the motor industry. The UNISURF System Proc. Roy. Soc. Lond. A 321, 207-218.
4. Bryson A., Ho Yu-Shi (1968). Applied Optimal Control. Optimization Estimation and Control. London, Blaisdell.
5. Kvasov B.I.. Yatsenko S.A. (1988). Solution of problems of isogeometric interpolation in the class of rational splines. Preprint of ITAM No. 3-88, Novosibirsk.
6. Kolchanov N.A., Anan'ko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignat'eva E.V., Goryachkovskaya T.N., Stepanenko I.L. (2000). Gene Networks. Mol. Biol. (Mosk.). 34:449-460.
7. Latypov A.F. (1974). A modification of the method of fastest descent. Izv. Sib. Otdel. Akad. Nauk SSSR, Ser. Tekhn. Nauk, 2(8).
8. Latypov A.F., Nikulichev Yu.V. (1985). Specialized group of optimization programs. Preprint ITAM SB AN USSR N15-85, Novosibirsk.
9. Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. (2001). A generalized chemical-kinetic method for modelling gene networks. Mol. Biol. (Mosk.). 35(6):1072-1079.
10. Ferguson J.C. (1964) Multivariable curve interpolation. J. ACM. II, 2, 221-228.

VIRTUAL REALITY AND REGULATORY SYSTEMS

Ratner V.A.

Institute of Cytology & Genetics, SB RAS, 630090, Novosibirsk, Russia
Novosibirsk State University, 630090, Novosibirsk, Russia

Key words: *regulatory systems, Virtual reality*

Resume

The **virtual worlds does not exist outside of regulatory systems or material carriers of information, connecting them.** In the sphere of Bioinformatics, as anywhere, there are three related groups of Regulatory Systems: Computers, Human Brains and Molecular Genetic Regulatory Systems. Each of them produced its own world of information images, that would be nominated as **Virtual World.** Our task was to confront the basic features of these Virtual Worlds with the features of the World of **Physical Reality**. The results of comparison seem to be surprising.

Introduction

The concept of Regulatory Systems is the central point of Theoretical Cybernetics (Von Neuman, Wiener, Lyapunov, Poletaev). The **Regulatory System** was determined as the system of production, storage, processing, realization, etc. of information. Information entities are the only products of regulatory systems. In the sphere of Bioinformatics, as anywhere, we meet three related groups of Regulatory Systems: Computers, Human Brains and Molecular Genetic Regulatory Systems. Each of them produced its own world of information images, that would be nominated as **Virtual World.** Our task will be to confront the basic features of these Virtual Worlds with the features of the World of **Physical Reality**.

Results

Our concept could be formulated as following. We postulate as axiom that **virtual worlds does not exist outside of regulatory systems or material carriers of information, connecting them.** Now we can indicate the basic features of the World of Physical reality and of known Virtual Worlds:

(A) There is the **Real World** of physical objects, events, systems, those have inherent material nature and are surrendered to natural laws.

(B) There is the **Virtual World of information entities of the human brain,** that could be produced by the humans (humanity) themselves, being realized or unrealized, in particular being dispersed in the process of teaching. The list of possible information products of the human brain is very long: (1) language, (2) personality, character, mentality, (3) mind, intelligence, (4) intuition, (5) soul, (6) genius, talent, (7) emotions, self-sensation, pain, joy among them, (8) knowledge, science, especially – mathematics, (9) faith, religion, including God, (10) picture of the Universe, World, (11) world outlook, ideology, (12) modelling, (13) tinkering, (14) creative work, painting, (15) music, (16) literature, poetry, (17) actorship, (18) fantastic actions, (19) dreams, (20) extrasensorship, (21) hypnosis, (22) mental disorders, hallucinations, (23) moral, ethics, (24) aesthetics, etc.

The laws and regularities of this virtual world are determined by the humans themselves. Virtual world in many respects is coordinated with the world of physical reality, but its entities being imitations of the real objects, are not necessary submit to the laws of the real nature (physics, chemistry, biology, etc.).

(C) There is the **Virtual World of computer information entities,** created by humans, adequate to virtual world of the human brains, and through them – to the real world. The laws of this virtual world are determined by the humans themselves – programmists, scientists, users, - and do not obligately coinside with the laws of physical reality.

(D) There is the **Virtual World of information entities of Molecular Genetic Regulatory Systems (MGRS) :** genes, functional sites, operons, genetic nets, genetic maps, mobile elements, catalitical activities, specificities, ontogenetic programs, genetic language, including the Genetic Code, etc. They are connected with the objects of real world through the inputs and outputs of MGRS. The laws of this virtual world were nobody determined, they are completely natural, and not completely arbitrary. They are restricted by the features of material carriers of genetic information (macromolecules and their systems).

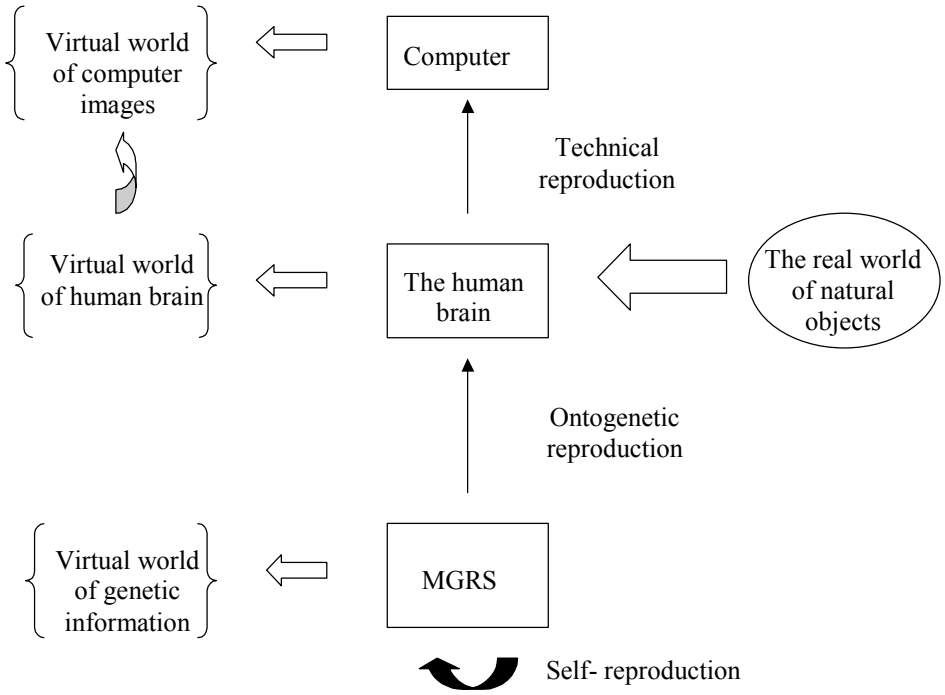


Fig. Relationship between the "worlds" and regulatory systems.

The place for the God

The human brain as regulatory system creates the virtual image of the surrounding world. In principle the Picture of the World, World Outlook must be adequate to the World of Physical Reality. Only then they will generally promote the survive of the individuals. However, they always obtain the personal characteristics, contain products of human creative work, fantasy, hypotheses, random features. Step by step this image became more objective, its fantastic or hypothetical features are substituted by information products of experience and science.

The Real World is very large, unbounded. Therefore, early or late in information picture of the world the problem appeared of the attitude to noncomparable, very large. They are: immense land space, desert, continent, ocean, typhoon, zunami, infinite sky above the head, Sun, Moon, stars, cosmos, lightning, etc. Furthermore, the humans meet so incomprehensible events as life and death, genius, hardly probable case, etc. Their appearance, behavior, presence are insuperable, incomparable with the scales of human brain. It is necessary to nominate them in the information picture of the world, to determine human's attitude to them. The humans could not quickly decide these problems, but they have no time to "stick" on this point. The experienced person (scientist) produce special scientific notion expressed this feature: infinity, eternity, case, combinatorics, very large, very small, etc. By other words, he postpones decision for tomorrow. The unexperienced person refuses to decide the problem, he acknowledges himself defeat.

It means that the **God has no place in real nature, but he has reserved place in the human information picture of the world**, no less, no more. The God and its belief, religion are information products of human brain. They have no any real power or authority apart of humans. But they could have very strong influence on the human behavior, decisions.

The any information entity could be recoded to different material carrier of information, regulatory system. It means that such personal information products of human brain as personality, soul, mentality, creativity could be dynamically supported on different computer, regulatory system without the limitations by time and place.

THE CONCEPT OF MOLECULAR GENETIC REGULATORY SYSTEMS (MGRS) AND POLYGENIC SYSTEMS

Ratner V.A., Vasylieva L.A.

Institute of Cytology & Genetics, SB RAS, 630090, Novosibirsk, Russia
Novosibirsk State University, 630090, Novosibirsk, Russia

Key words: *MGRS, polygenic systems, genetic nets, limiting genes*

Resume

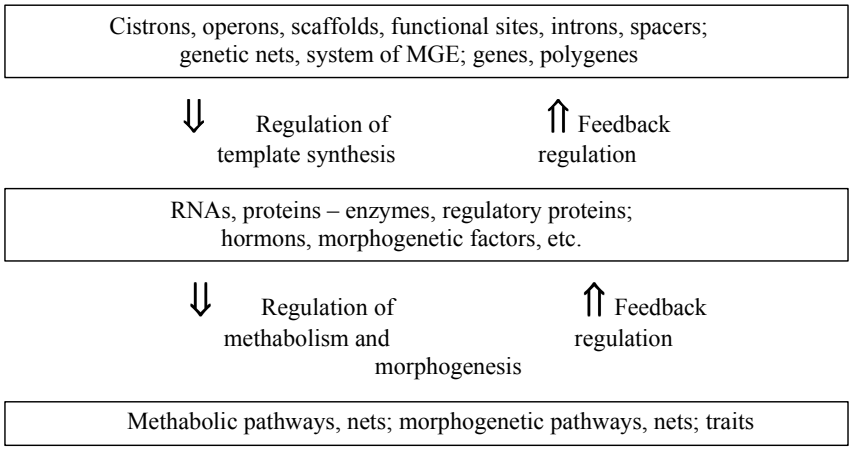
All attributes of polygenic systems – traits, limiting genes (Mendelian genes, oligogenes), nonlimiting genes (polygenes) – could be extracted from genetic, metabolic and morphogenetic nets. The transition from description by nets to description by polygenic systems of quantitative traits is actually extraction of non-limiting genes and linearisation of their action to trait.

Introduction

The concept of MGRS was developed earlier by us, being now the methodological ground of the description of genetic nets. This concept contains some principles, used for description of complex MGRSs: block-modular principle, principles of limiting factors, etc. The concept of polygenic (QTL) systems had appeared in classical genetics for description of complex genetic determination of quantitative traits. Now it is necessary to find conformity of the key features of these concepts.

General Results

1. The MGRSs are the complex systems. Below there are the estimates of gene numbers based on the results of complete sequenation of genomes:
E. coli ~ 4900, S. cerevisiae ~6000, D.melanogaster ~ 13600, H.sapiens ~ 38600.
We know the gene types from molecular-genetic analysis. The principles of their regulation in procaryotes and eucaryotes we know in general view, though not up to the end.
The block-modular scheme of the MGRS construction of eucaryotic organism is shown on the Fig. The modules were extracted in accordance with the discussed problem.



2. **The genes, polygenes** – product of classical genetic analysis through traits. Any identified feature, structure could be used as a trait. What is the determination of the trait in terms of complex MGRS?
a) Any trait in the net (genetic, methabolic, morphogenetic) has its **nearest environs** by the net. These are elements those by net are either directly participated in interactions, or are shortly distant from there.The environ is important but not exhaustive component of the trait determination, but no one action for the trait escapes the environ.
b) **The principles of limitation**. The position of limiting link is determined by the structure of the net, and especially of the nearest environ of the trait. As the limiting factors (genes, enzymes, etc.) there were nominated the components ,

whose changes (mutations, regulatory changes) were essentially expressed in the trait variations. As example, such are auxotrophic mutations in metabolic pathways. In classic genetic terms, the limiting genes are actually **Mendelian** ones. The other genes controlling the elements of the nearest environ of the trait, are **polygenes**. They don't limit expression alone, but could be expressed in total through the limiting Mendelian gene. They could become limiting by mutations or by regulation. The regulation is the most effective being introduced through the limiting links.

3. The **mathematical description** of the nets – nonlinear differential equations of biochemical kinetics, where the rules of the trait formation are determined only by the structure of the net. The mathematical description of polygenic systems – nonlinear equations of polulation dynamics, where the rules of trait formation are usually additive. It could be expected that actually the linear sums of the small contributions of polygenes to trait are the results of linear Taylor decomposition in the environ of stable or stationar points.

4. So, in genetic, methabolic and morphogenetic nets it is possible to extract all the attributes of polygenic systems: traits, limiting genes (Mendelian genes, oligogenes), nonlimiting genes (polygenes). Actually, the transition from the description of the nets to description of polygene systems of the quantitative traits is based on the extraction of nonlimiting genes and linearisation of their contributions to trait.

Acknowledgements

This work was partially supported by Russian Foundation of Basic Researches (RFBR) grant № 00-04-49499, and by grant of the Ministry of Education of RF "Russian universities – fundamental researches" (№ 1760).

References

1. Ratner et al. (1996). Molecular Evolution. Springer-Verlag, Berlin e.a.
2. Ratner V.A., Yudanin A.Ya. (1999). Population dynamics of an Additive Polygenic System during Limiting Selection. Genetika *(*Russ.). 35, 6, 853-861.

THE CONCRETE POLYGENIC SYSTEM *RADIUS INCOMPLETUS* AS EXAMPLE OF ACCORDANCE OF THE MGRS NETS AND POLYGENIC SYSTEM

Ratner V.A., Vasylieva L.A.

Institute of Gytology and Genetics, SB RAS, 630090, Novosibirsk, Russia
Novosibirsk State University, 630090, Novosibirsk, Russia

Key words: *genetic and other nets, polygenic system,* radius incompletus*, molecular mechanisms, genomic MGE system, interaction of polygenes and MGE*

Resume: Now we can indicate some molecular equivalents to Mendelian genes and polygenes, elements of genetic net, to estimate the complexity and topography of the polygenic system *radius incompletus*, to estimate the role of the genomic MGE system in interaction with polygenic system of *ri.*

Results

I. a) The normal polygenic system *radius incompletus (ri)* is responsible for topography and formation of drosophila wing radial vein (L2). Mutation *radius incompletus* brakes the vein into two fragments, their lengths being quantitative traits under the control. From the qualitative view point, this mutation is Mendelian.
b) Let us take into account two features, characteristic for drosophila.
- According to Waddington, the genetic systems controlling the formation of morphological traits must be subdivided for two subsystems: **subsystem of prestructure,** that controls the topography of the trait, its plan, and **subsystem of execution**, that directly participates in the formation of the trait.
- Drosophila has several stages of development. The wing radial vein is physically formated on the puppal stage, but the regulatory events, connected with topography of the vein, are realized on the larvae stage in the specific imaginal discs.
c) The molecular nature of oligogene *ri* was revealed by the group of German and American scientists (Linde et al., 1998). Two neighboring genes are located in segment 77E1 of Bridges cytological map: *knirps (kni)* and *knirps-related (knrl)*. They code the factors of transciption in superfamily of receptors of steroid hormones. Mutation *radius incompletus* is regulatory mutation in the locus *kni/knrl,* these loci being nonexpressed in many cells usually predetermined to participate in formation of the vein. Physically, this mutation is deletion in restriction fragment of the length 1.7 kb, located upstream of the gene *kni* by 2.5 kb.
d) The genes *kni/knrl* are the elements of the subsystem of prestructure, active yet in the imaginal discs of larvae stage. They have 4 regulatory functions:
1) They induce the expression of executing genes of the puppal stage (*rho* – downstream of *kni/knrl).* From the other side, they are stimulated by regulatory gene *salm* (upstream of *kni*). Its short signal *X* induces loci *kni/knrl* in the cells along the forward border of the domain of expression in imaginal disc.
2) They repress the formation of the vein in neighboring cells (vein fate).
3) They support their own expression by autoregulation feedback.
4) They sharpen the forward border of imaginal zone by the negative feedback.
Thus, the genes of the subsystem of prestructure obtain the regulatory function and act on the larvae stage in imaginal discs. The genes of the subsystem of execution, to the point participating in formation of all different wing veins, are functioning on the puppal stage. It is real to investigate the **genetic net** of formation of the wing radial vein.

II. The polygenic system *radius incompletus*

The information about **polygenes** is substantially less, in particular their molecular nature is unknown, though there are some general results and ideas.
a) There are no less than 15-20 polygenes, located along all the genome. There are polygenes of small and large effect. The expression of the trait could be essentially changed by the change of the temperature of cultivation and by selection. Apparently, the temperature could influence through the kinetics of morphogenetic processes, and selection – by selection of alleles of polygenes.
b) The step-wise temperature change ($29^o \Rightarrow 18^oC$) is effective only on the puppal stage, i.e. when the prestructure is already determined, and the subsystem of execution is functioning. The radial vein is formed from two ends. Kinetics of the formation could be different, producing the temperature changes of the trait. There is possible the movement of the cells with determined prestructure from imaginal discs to their locations on the wing. In this case there may be different in the rates.

c) The selection in isogenic line is noneffective, it means that the genetic variability of polygenes is absent. Selection in heterogeneous lines, mixtures of lines, induced lines is effective. Hence there were or there appeared the variability of polygenes.
d) The versions of hypotheses about molecular nature of polygenes (all – the cistrons):
1) Mendelian genes (nonlimiting) of different systems, those express some influence on expression of the trait by the nets. Mainly these are the genes from the nearest environ of the trait.
2) Regulatory genes with threshold action, that could activate transcription.
3) Special small ORF-s, modifying activity of Mendelian genes. There is precedent of such situation: the genome of phage T4 containes 150 small ORFs, repressing the expression of *E.coli* genome. In the programs of complete sequenation of genomes the small ORFs were usually omitted from consideration. May be we lose this information?
4) Genes of common cell transcription activators under their action on the mutant (limiting) Mendelian genes.

III. The participation of the MGE system in expression, variability and different features of the polygene systems

a) The features of Mobile genetic elements (MGE)
- The MGE copies inserted mainly to the regulatory zones of Mendelian genes.
- MGE contained the motifs of different transcription sites: enhancers, silencers, HS-sites, etc.
- MGE copies are dispersed along the genome, drosophila contains tentatively 1 MGE copy per 2-3 genes.
- There are direct experimental facts of regulatory reorganization of Mendelian genes after MGE insertions to their regulatory zones.
- MGE can perceive the external factors (HS, different stress factors, ethanol, etc) as the signals of transcription and transposition.
- For insertion of MGE copy near some locus it is necessary its "unpacking" in chromatin (scaffold).
- For expression of regulatory influence, activation of gene, visible as the response to selection, it is necessary its "unpacking" in chromatin (scaffold).
b) Molecular mechanisms of induction of transpositions (transcription).
- The genomic system of response to HS, induced the transcription of retrotransposons through the functional sites of HS.
- Different inducible systems (poisons, heavy metals, detergents, ethanol, etc.) acting perhaps by the same HS system.
- The system of repair of two-stranded DNA brakes with their curation by MGE copies. Two-stranded brakes appeared after treatment by γ-irradiation and different strong factors.
The genetic treatments (isogenization, inbreeding), could induce transpositions by the same mechanisms of HS-response. The mechanisms of action of outbreeding are unknown, but it is possible the role of heterozygosity by large chromosome segments, prevented the synapsys (as example – by large inversions in balancer chromosomes in the course of isogenization).
c) Molecular population mechanisms of the MGE patterns response to selection by quantitative trait.
- MGE copies are subdivided in 3 groups: independent copies, markers and modifiers of polygenes.
- MGE-modifiers, activators of polygenes expressed the fast selective response together with polygenes.
- MGE-markers, independent copies and all the rest genome expressed the fast unselective response, by the mechanism of strong selective inbreeding under the truncation selection and high progeny production.

IV. Sum total

a) In the complex MGRSs it is possible to describe the **limiting** genes as Mendelian, and **nonlimiting** – as polygenes (QTLs).
b) It is possible to describe and investigate the **genetic net** of the formation of wing radial vein.
c) MGEs compose the **all-genomic system** capable to regulatory reorganization of the genome, especially of ontogenetic subsystems, and to response for external and genetic stresses.
d) The MGE copies could be considered as "movable cassettes of functional sites".
e) The system of MGEs is expected to play the important role in evolution of polulations by the strong induction of genetic variability and fast response to selection

Acknowledgements

This work was partially supported by Russian Foundation of Basic Researches (RFBR) grant № 00-04-49499, and by grant of the Ministry of Education of RF "Russian universities – fundamental researches" (№ 1760).

References

1. Lunde K. et al. (1998) The *knirps* and *knirps-related* genes organize development of the second wing vein in drosophila. Development. 125, 4145-4154.
2. Kutter E. (1996). Analysis of Bacteriophage T4 based on the completed DNA sequence. Integrative Approches to Molecular Biology (eds. J.Collado-Vides, B.Magasanik, T.F.Smith). The MIT Press, Cambridge, Mass. – London, England. 2, 13-28.
3. Ratner V.A., Vasylieva L.A. Mobile genetic elements (MGE): "selfish DNA" vs. functional elements of the genome. The Modern Problems of the Evolutionary Genetics (Eds. V.K.Shumny, A.L.Markel), Novosibirsk, ICG SB RAS. 145-170.

Addendum: The last confirmation of the concep: Induction of the MGE transpositions by steams of ethanol

[1]Lopukhova E.D., [1]Antonenko O.V., [1, 2]Vasilyeva L.A., [1, 2]Ratner V.A., [1]Bubenshchikova E.V.

The patterns of MGE 412 was revealed by method of *in situ* hybridization of the probe, containing the copy of this MGE, with DNA of polytene chromosomes of drosophila larvae salivary gland cells. On the basis of *in situ* hybridization in F1 larvae transpositions were found in 10 sites of genome. After treatment, the average transposition rate increased by 1-2 order of magnitude (compared to control). This basic induction level was observed at various doses of ethanol.

In F1 larvae transposition were found in 11 sites of genome: 12B, 34B, 36B, 42B, 43B, 56B, 56B, 56E, 75C, 97DE and 100B. This sites were not present at a starting isogenic line, hence, their display were induced by steams of ethanol on cells of generative lines of males during spermatogenesis.

As was already mentioned, series of works about induction of transpositions of the MGEs by stress factors in different isogenic lines of drosophila were earlier executed. Thermal influences (isogenic lines № 2-2, № 16, № 51) and γ-irradiation (isogenic line № 49) were among this factors. The analysis has shown, that MGEs "prefer" to insert in the certain sites, so-called "hot sites" (34B, 43B, 97DE and others). This feature can be revealed in new experiment with ethanol also. The greatest number of insertions were found in sites 56E - 3, 97DE - 4 and 100B - 4 times, i.e. 11 from 23. At augmentation of sample, probably, it will be possible to name the site 97DE "hot" for all types of treatment, including processing by steams of ethanol.

The fact is that as a result of different ways of stressful influence on a genome, its reaction was mostly the same.

The average rate of ethanol steams induced transpositions is 4.8×10^{-2} events per site per spermium per generation, that is 1-2 order of magnitude more than of spontaneous transpositions. Hence, the phenomenon of the induction of transpositions of the MGEs with steams of ethanol can be read proved.

FORMAL DESCRIPTION OF THE TREMATODE ECOPARASITIC SYSTEM ON USING THE GENENET DATA FORMAT

Vodyanitskii S.N., Yurlova N.I. [1]***, Suslov V.V.*** *

*Corresponding author
Institute of Cytology and Genetics, SB RAS, Novosibirsk, 630090, Russia
[1] Institute of Animal Systematics and Ecology, SB RAS, Novosibirsk, 930091, Russia

Key words: *databases, GeneNet, format, trematodes, ecosystems*

Summary

Motivation: The set of databases where the experimental and field information on biodiversity is annotated on the basis of common ontology in compatible formats would allow (1) application of a unified program toolkit for data analysis and modelling and (2) analysis and modelling of biosystem operation at different hierarchical levels, from a gene to an ecosystem.
Results: The format of GeneNet database has been optimized with regard to annotating information and portrait-based visualization of network interactions in ecosystems in terms of hierarchical levels. An example is the portrait-based description of the life cycle of the trematode *Echinoparyphium aconiatum.*
Availability: http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet

Introduction

Processes occurring in organisms and ecosystems can be formally described as circulation of information, matter, and energy within a network. The nodes of such a network in an organism are genes and/or proteins; in an ecosystem, species and/or living forms. They are linked with regulatory interactions determining transformation of information, matter, and energy. The regulatory interactions can be divided into (1) *reactions*, in which interaction of several entities gives rise to a new entity or one of the old entities disappears or changes, and (2) *regulatory effects* affecting *reactions* (Ananko et al., 1998). This formal consideration forms the grounds for development of a unified computer technology allowing description of interactions within biological systems in a single format throughout all the hierarchical levels, from a gene to an ecosystem. This technology can be used in Internet-available biodiversity databases on biodiversity or added to the toolkit used in information biology for modelling ecosystem processes. A unique database GeneNet (Ananko et al., 1998) was developed at the Laboratory of Theoretical Genetics, Institute of Cytology and Genetics, Novosibirsk, for description of the structure of gene networks. Extension of the format of this database is the first stage in development of this technology. We have chosen the parasite–host system for a pilot description for the two following reasons: (1) this system is well studied and (2) its system-forming trait, living cycle, is readily identifiable, as opposed to ecosystems of free-living organisms, which facilitates its formal consideration.

Methods

Data on the structure of network interactions in the course of the complex life cycle of the trematode *Echinoparyphium aconiatum* were accumulated in the GeneNet database (Ananko et al., 2002) with the use of interactive data input through Internet (Ananko et al., 1999). Information was obtained by annotating scientific publications.

Results

The object of formalization. *Echinoparyphium aconiatum* has a typical trixenic (three-host) life cycle. The first and second bridging hosts are freshwater gastropods, and the definitive host is waterfowl. Hermaphroditic maritae, parasitizing in birds, lay eggs. The eggs release freely mobile non-eating miracidia. During their life, they must find a mollusk and invade it. The mollusk becomes their first bridging host. A series of metamorphoses (miracidium→sporocyst→maternal redia) is followed by parthenogenesis of rediae and birth of a cercaria hemipopulation. The free-mobile non-eating cercariae encyst in the second bridging host forming metacercariae. They are eaten together with the second host by the definitive host, where they become maritae. At a low ambient temperature or other unfavorable seasonal factors, cercariae may encyst in the first bridging host, which, thereby, plays also the role of the second one.
Basics of formalization. All components of the life-cycle network were divided into *elementary objects, linkages (elementary events),* and *"processes"* (any developed events regarded as a whole, their internal structure not being taken into consideration). The *elementary objects* included (1) all structure–function compartments of the networks from a gene to host organisms and (2) stages of the parasite development, regarding different steps of ontogenetic stages as different objects. For example, a free-swimming miracidium and a miracidium that had penetrated into the host were considered to

be different objects (Fig. 1A); (3) high/low molecular weight compounds with various functions (signal, trophic, immune, etc.) released or consumed during the life cycle (Fig. 2B); (4) environment (air or water); (5) general environmental factors (temperature, insolation, pH, etc.) (Fig. 1B). The *linkages* were subdivided into (1) reactions, that is, the events giving rise to new components of the network (for example, the release of a miracidium→searching movements of the miracidium) and (2) regulatory events (response switch-on, switch-off, enhancement, or inhibition). "*Processes*" are convenient if details of a phenomenon are either unknown or needless. In the detailed scheme, "*processes*" include (1) metabolic, growth, morphological, and other processes in a host/parasite organism; (2) behavioral processes; (3) integral traits related to the life cycle of trematodes (for example, "feeding" (of the second bridging host by the definitive host), "differential death-rate", etc.); and (4) deviations of the general environmental factors from the optimum, for instance, pH or temperature fluctuations (Fig. 1). In the simple scheme, "*processes*" denote stages of the life cycle of the parasites, as they are regarded as a whole (Fig. 2). In the detailed scheme, life-cycle stages are *elementary objects*, where ontogenetcal events resulting in metamorphosis and change of generation can be represented as "*processes*" and then, with accumulation of new data, as gene networks.

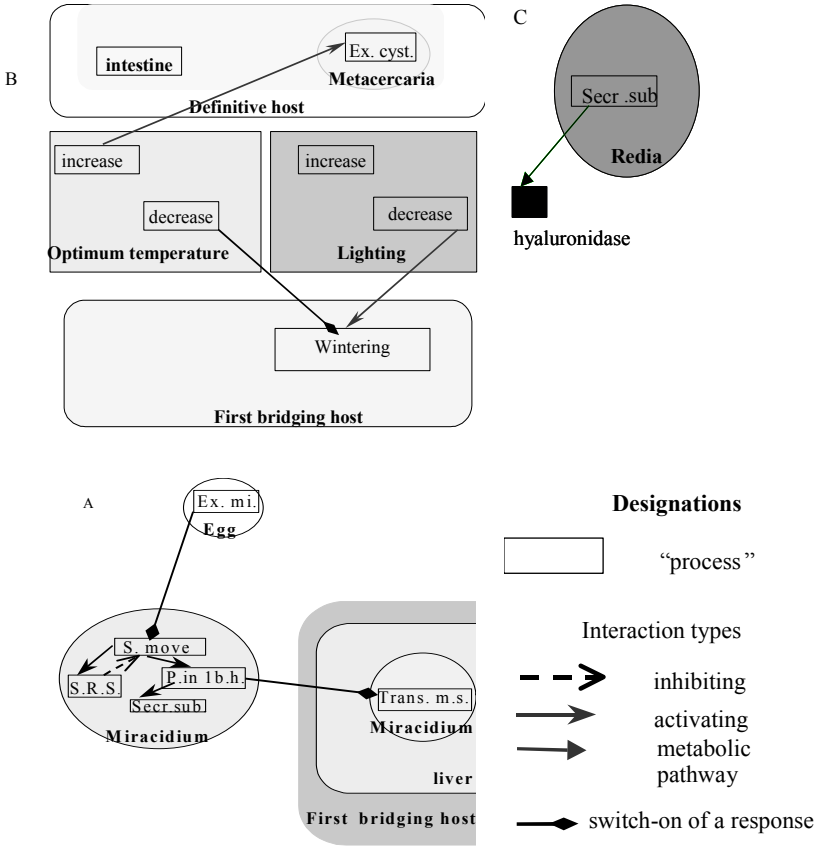


Fig. 1. Examples of event visualization in the detailed network of the ecoparasitic system (in a simple form): (A) Release of a miracidium from an egg and percolation into the bridging host; (B) General environmental factors: an example of influence on "processes"; and (C) Hyaluronidase production by rediae. Abbreviations of "processes": **Ex mi**, exit of miracidium from an egg; **S. move**, searching movements of a miracidium; **S.R.S.** storage substance spending; **Secr. sub.**, secretion of substances; **P. in.1b.h**, penetration into the 1st bridging host; **Trans.m.s.**, transformation into a maternal sporocyst; and **Ex.cyst**, exit from a cyst.

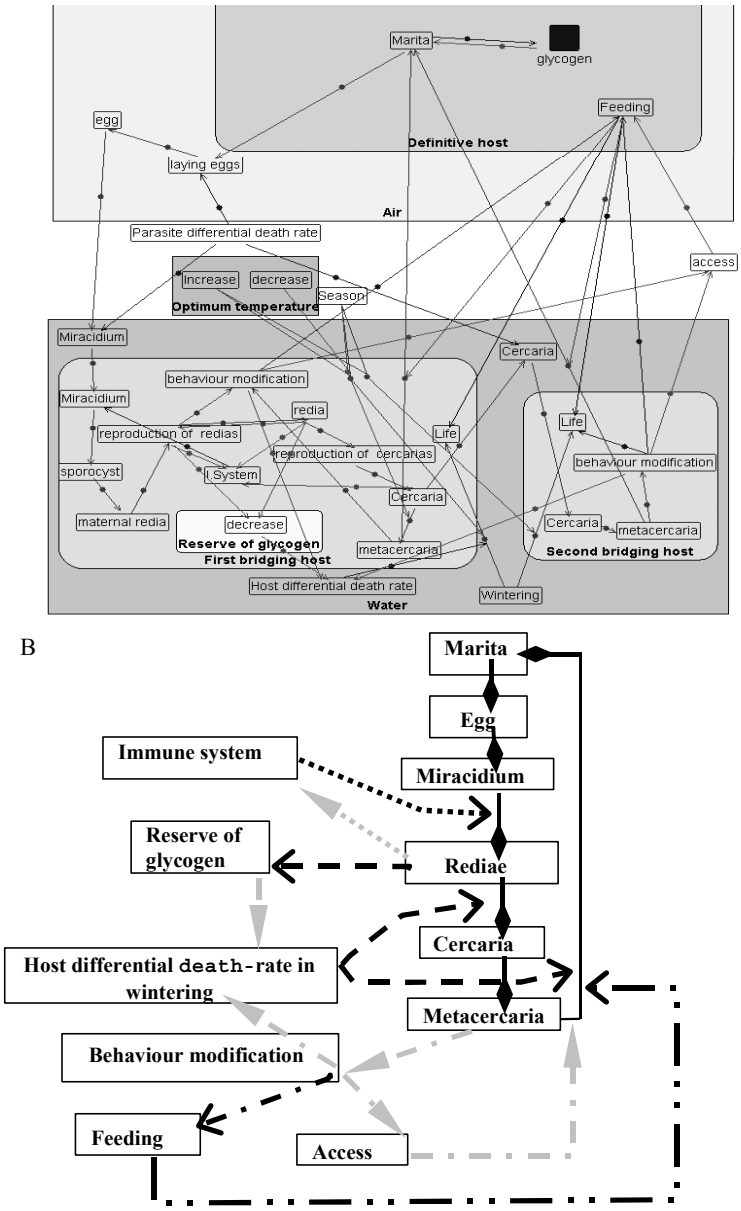


Fig. 2. (a) A simplified network of the trematode ecoparasitic system and (b) its feedback loops.

In **a,** life cycle stages are regarded as *processes*. The *processes* feeding, access, and I system denote, respectively, eating of a bridge host with a definitive host, accessibility (of the second bridge host as food for the definitive host), and the immune system.

Designations for **b:** dark arrow ( ), inhibition; light arrow ( ), activation, diamond-headed arrow ( ), life cycle of trematodes; ••••••••, the negative feedback loop operating at the level of the organism of the first bridge host; – – – –, ditto, at the level of the phenocycle for the first bridge host; -•-•-•-•-, ditto, for the second bridge host; -••-••-••-; ditto, at the level of trophic linkages between the first and second bridge hosts.

Discussion

The network of the ecoparasitic trematode system is cyclogram with one subcycle corresponding to redia parthenogenesis (Fig. 2A). The growth of hemipopulations of free-swimming stages is compensated by their differential death-rate. Three negative and one positive feedback loops can be recognized in the network. The negative-feedback loops operate at various hierarchical levels: the first loop, at the level of the organism of the first bridging host; the second, at the level of its phenocycle; and the third, at the level of trophic linkages between the bridging and definitive hosts (where the positive feedback loop operates too). After penetrating the first bridging host and starting parthenogenesis, the parasite, being an antigen source, activates the host immune system. As a result, a miracidium subsequently invading the host encounters a powerful immune response and, as a rule, dies. This is one of the mechanisms limiting the intensity of invasion (Fig. 2B). Another mechanism discards strong invaded mollusks. They cannot accumulate enough glycogen. Moreover, their behavior changes (especially, if the mollusk is the first and second host simultaneously). The mollusk either remains on the silt surface or does not dig in deep enough. Therefore, strong invaded mollusks die under adverse environmental conditions and, as a rule, fail during wintering (Yurlova, 2000; Fig. 2B). The latter mechanism concerns the food behavior of birds. Strong invaded mollusks lying on the silt surface are more readily accessible than healthy, lying deep. Thus, a bird can follow one of two food strategys: either reduce spending energy by picking mollusks from the silt surface or reduce the risk of invasion by digging mollusks from silt and spending more energy. The former strategy implements the positive feedback loop; and the latter, the negative one. It is worth noting that birds avoid eating large mollusks lying at the bottom surface (Yurlova, 2000; Fig. 2B). Thus, we observe an apparent similarity between the gene network of an

organism, supporting homeostasis, and the ecoparasitic system supraorganism network. Both of them are controlled at various hierarchical levels by negative feedback loops. Further information would reveal more regulatory loops ensuring homeostasis of this system.

Conclusions

1. A portrait-based description of the life cycle of the trematode *Echinoparyphium aconiatum*, a supraorganism ecoparasitic system, was entered to a database with the use of a format developed for describing gene regulation mechanisms and other molecular biological pathways of information transfer in an eukaryotic organism.
2. The format allows consideration of interactions between the parasite and the host at all hierarchical levels.
3. The block structure of the format allows easy extension of the network by introducing new hierarchical levels (interparasitic, cellular, molecular-genetic, etc.), as new information is accumulated.
4. Investigation of the ecoparasitic network system demonstrates its similarity to the gene network of an organism. Thus, the experience and tools developed for modelling gene networks can be used for simulating the network of an ecoparasitic system. By now, the first step has been made: Some network interactions in the complex life cycle of the trematode *Echinoparyphium aconiatum* are described in a GeneNet-like format. Mathematical modelling will be the next step.

Acknowledgements

The study was supported in part by the Russian Foundation for Basic Research (grants № 01-07-90084 and 02-07-90359); Russian Ministry of Industry, Science, and Technologies (grant № 43.073.1.1.1501); and Siberian Branch of the Russian Academy of Sciences (Integration project № 66). The authors are grateful to V.V.Gulevich and G.B.Chirikova for translating the manuscript into English.

References

1. Ananko E.A., Kolesov G.B., Kolpakov F.A., Kolchanov N.A. (1998). GeneNet: a database for gene networks and its automated visualization. Bioinformaticsю 14, 529-537.
2. Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. (2002). GeneNet: a database on structure and functional organization of gene networks. Nucl. Acids Res. 30**,** 398-401.
3. Ananko E.A., Kolpakov F.A. (1999). Interactive data input into the GeneNet database. Bioinformatics. 15, 713-714.
4. Yurlova N.I. (2000). Change in mollusk behavior related to trematode invasion. In: Problems of hydroecology at the century borderline [in Russian], St.-Petersburg.