RUSSIAN ACADEMY OF SCIENCES
SIBERIAN BRANCH

INSTITUTE OF CYTOLOGY AND GENETICS
LABORATORY OF THEORETICAL GENETICS

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

Volume 2

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

IC&G, Novosibirsk, 2004

International Program Committee

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

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

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

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

Philipp Bucher, Swiss Institute for Experimental Cancer Research, Switzerland

Gennady Erokhin, Ugra Research Institute of Information Technologies,

Khanty-Mansiysk, Russia

Jim Fickett, AstraZeneca, Boston, USA

Mikhail Gelfand, GosNIIGenetika, Moscow, Russia

Sergey Goncharov, Sobolev Institute of Mathematics, Novosibirsk, Russia

Igor Goryanin, GlaxoSmithKline, UK

Charlie Hodgman, GlaxoSmithKline, UK

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

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

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

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

Nilkolay Podkolodny Institute of Cytology and Genetics, Novosibirsk, Russia

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

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

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

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

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

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

Eugene Zabarovsky, Karolinska Institute, Stockholm, Sweden

Lev Zhivotovsky, Institute of General Genetics, Moscow, Russia

Local Organizing Committee

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

Anatoly Kushnir, Institute of Cytology and Genetics, Novosibirsk

Natalia Sournina, Institute of Cytology and Genetics, Novosibirsk

Galina Kiseleva, Institute of Cytology and Genetics, Novosibirsk

Katerina Denisova, Institute of Cytology and Genetics, Novosibirsk

Andrey Kharkevich, Institute of Cytology and Genetics, Novosibirsk

Yuri Orlov, Institute of Cytology and Genetics, Novosibirsk

Organizers

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

Siberian Branch of the Russian Academy of Sciences

All - Russian Society for Geneticists and Breeders

Ugra Research Institute of Information Technologies

INTAS

Sponsors

Russian Foundation for Basic Research

AstraZeneca, Boston, USA

Information Sponsors

Biophysics (Russian)

In Silico Biology

Contents

COMPUTATIONAL SYSTEMIC BIOLOGY

COMPUTATIONAL EVOLUTIONARY BIOLOGY

NEW APPROACHES TO ANALYSIS OF BIOMOLECULAR DATA AND PROCESSES

BIOINFORMATICS AND EDUCATION

Introduction

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

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

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

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

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

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

The huge volume of experimental data that has been acquired on genome structure, functioning and gene expression regulation demonstrate the blistering growth. Onrush of the volumes of experimental data is observed in the recent years due to the fact that genome deciphering became a technical routine task produced at high speed. Unprecedented large bulk of experimental data emerge under studying of molecular-genetic systems and processes with application of microarray analysis technique. Unwrapping of large scale studying in proteomics is accompanied by accumulation of very large information pools on primary and spatial structures and functioning patterns. That is why development of informational-computational technologies of novel generation is a challenging problem of contemporary bioinformatics. Bioinformatics has entered that very phase of development, when decisions of the challenging problems determine the realization of large-scale experimental research projects directed to studying genome structure, function, and evolution. Essentially that bioinformatics is recognized as a necessary element for modern experimental research. It is widely used as at the stage of experimental designing and data interpretation, as for solving fundamental problems of organization and evolution of molecular-genetic systems and processes. By analyzing the papers submitted for publication in the two-volume issues of the BGRS' 2004, the Program Committee came to a conclusion that participants of the Conference have concentrated their attention at consideration of the hottest items in bioinformatics listed below:

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

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

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

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

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

BGRS
2004

COMPUTATIONAL SYSTEMIC BIOLOGY

A SYSTEMIC APPROACH TO COMPLEX, MULTI-FACTOR AUTOIMMUNE DISEASES AIMED AT CREATION OF ADEQUATE MODELS OF PATHOLOGIES (e.g. PSORIASIS)

Abdeev R.M.

Vavilov Institute of General Genetics RAS, Moscow, Russia; Center for Theoretical Problems of Physical-Chemical Bases of Pharmacology RAS, Moscow, Russia, e-mail: abdeev@vigg.ru

Keywords: *psoriasis, autoimmune diseases, model of pathologies*

Summary

Motivation: The fact that an extremely large body of experimental data has been accumulated, theoretical understanding of some of the immune system units has been achieved and a large amount of new, more detailed and specific information has become available, allows one to approach structural and functional modeling of the immune system under normal and pathological conditions.

Results: In this work we present a proposed networking character of interactions between immune system components at the cellular level and at the level of soluble mediators of inflammation (the cytokine network) under psoriasis disease. Results of this study form a basis for computer modeling of a pathological condition.

Introduction

The immune system of man is one of the most complex functional systems of the body. The elements of acquired immunity with all their intricate functional links indeed represent a unique system that implements the principles of stochastic, programmed, and programmed/adaptive control. Many systemic diseases such as malignant malformations, autoimmune diseases (e.g. psoriasis, diabetes, and multiple sclerosis), immunity deficiencies and many others are in fact various pathological states of the immune system. The creation of an adequate structural-functional model of the immune system is deemed to be a most urgent and, at the same time, most difficult challenge for systemic computer biology, second only perhaps to modeling of the higher nervous system. To simplify the model, an attempt was made to represent the immune system in an hierarchical way and to consider processes that take place in it separately, both within a hierarchy level and within each separate subsystem of each of the levels. Once the properties and behavior of individual subsystems are modeled, it would be possible to reassemble the whole system as an aggregate of the subsystems. Psoriasis was chosen as an example modeling of a complex multi-factor autoimmune pathology.

Model

The immune system consists of two main subsystems, the inherent (unspecific) immunity and the acquired (specific) immunity (Fig. 1). The inherent immunity, characterized by a rapid immune response, is most effective against unspecific bacteria and microorganisms.

This is why to protect the body against attacks of such pathogens the acquired immunity system is switched on, its function being to detect microorganisms inaccessible to the inherent immunity system and to generate signals for its targeted activation.

Hence, in a very general way, the immune system can be envisaged as consisting of two hierarchical levels, the unspecific immunity being the principal executor in protecting the body from microbes. The principal executor of inherent immunity is, in turn, the system of complement which is activated automatically upon infection of the body with unspecific bacteria (the C3 pathway). A cascade of reaction takes place, resulting in a rapid eradication of the infection owing to both functioning of complement proteins and attraction to the infection area of macrophages, the principal immune system's executors at the cellular level.

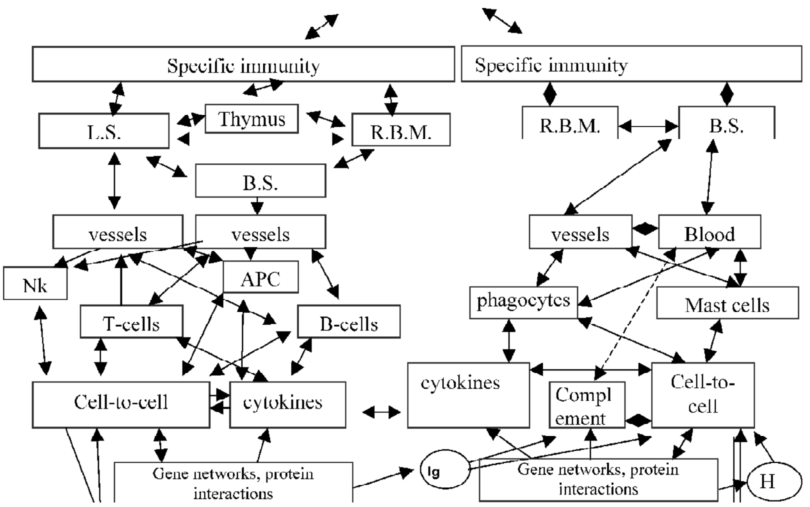


Fig. 1. Hierarchical model of immune system. L.S. – lymphatic system, R.B.M. – red bone marrow, B.S. – blood system, Nk – normal killers, APC – antigen presented cells, Ig – immunoglobulins, H – histamine.

When the complement system fails to detect a foreign microorganism, the specific immunity becomes activated. B-cells in this case start to produce specific antibodies, which can not eradicate the infection but can form aggregates with foreign polypeptides or cells. Such aggregates can be detected by both macrophages and the complement system proteins. As a result, this again leads to activation of the C3 pathway of complement, followed by the reaction cascade described above. In order to start efficiently producing antibodies, B-cells in turn should receive an activator signal from their hierarchical superiors, T-cells, which are at the top of the cellular immunity hierarchy. Unlike other cells of the immune system, T-cells have body's own cells as their targets. T-cells are produced in red bone marrow and then migrate to the thymus where they receive "training" whose main objective is selection for a strict discrimination between "own" and "foreign" and for adequate and strictly specific interaction with other cells of the body. T-cells are subdivided into cytotoxic T-cells (Tc) whose functions consist mainly in liquidation of body's own infected cells, and T-helpers (Th). Depending on the spectrum of regulatory cytokines produced, two types of T-helpers, Th1 and Th2 can be distinguished, although this subdivision is rather arbitrary. Th2 are most effective in specifically stimulating and activating B-cells, while Th1 are mainly involved in local inflammation and in autoimmune reactions.

The T-cells have at their disposal a universal regulatory machinery with a great potential – numerous soluble regulatory mediators (the cytokines produced by T-cells in certain quantities and at proper times) and numerous specific cellular receptors. Therefore, the cellular response and the regulatory signal are the results of very fine and subtle interactions, through cell-to-cell contacts and through cytokine "words", where each cytokine is just a letter in a word (or even in a phrase). Such a mechanism allows the immune system to encode information at the receptor, cellular and inter-cellular levels.

Construction of a hierarchical model of the immune system is based on the following principles. Since each complex system of the body has certain inherent frequencies of its internal processes, it is postulated that the higher the frequency the lower the hierarchy level of the subsystem, and that more fundamental subsystems also correspond to a lower hierarchy level. The existence of higher hierarchy levels imparts greater stability to the system as a whole. On the basis of these postulates, and also of the principles and types of regulations existing in the body, the immune system occupies the next hierarchy level after the higher nervous system. Then, down the hierarchy order go the nervous, endocrine and blood systems, each of which consists of the organs and tissues, cells, cell structures, complex intermolecular interactions (gene networks, etc), simple inter-molecule interactions, genetic molecules (proteins, nucleic acids) and simple molecules, the metabolites. Such a concept allows one to structurally define the whole system and to distinguish hierarchy levels, which is convenient for structural and functional modeling.

In describing complex living systems one always comes across the fundamental question of causes and effects. There is one paradigm which states that since genes control the organism's intricate functions, it is most appropriate to make emphasis on study of genome functioning as it determines the body's behavior: to construct gene networks, establish most important regulatory components among them, etc (Kolchanov *et al*., 2000). Such an approach has already proved effective for the modeling of various genetic processes at the molecular level. The other paradigm states that complex biological systems employ genes as universal tools to perform various functions, i.e. that complex living systems control the genes. If this latter paradigm is taken, it would seem most effective to study the hierarchy of system interaction, and the gene networks as one of the hierarchy levels. There is an increasing body of evidence suggesting that epigenesis in a broad sense of the word play an important role in inheritance of biological information, thus speaking in favor of the latter paradigm. Solving the conflict between the two paradigms, which is based entirely on cause-effect relations, can only be done at the philosophic level but not by means of experimental scientific endeavor. However, if one takes the view that the genes and their superstructures (the complex systems) constitute a whole, and that at times the control may be handed over to either ones or the others, then the input of cause-effect relations would be localized, thereby making the modeling all the more convenient.

We believe that the methodology of construction of gene networks may be applied to other hierarchy levels as well at each of which, according to our postulates, the networking structure of interactions is predominant. As an example, we shall discuss below the networking nature of interactions between individual subsystems of human specific immunity, a disturbance in which leads to psoriasis.

Results and Discussion

Psoriasis is presently regarded as a dermatological disease characterized primarily by hyper-proliferation of keratinocytes, their improper differentiation, and alterations of the cell cycle and apoptosis system in epidermal keratinocytes resulting from a misbalance in the specific immune response systems. It is believed that Th cells play an important functional role in inducing inflammation, and the nature of the pathology allows classifying psoriasis as an autoimmune disease. As a systemic disease, psoriasis may be studied at numerous hierarchy levels of the organism. In this work we present a proposed networking character of interactions between immune system components at the cellular level and at the level of soluble mediators of inflammation (the cytokine network).

The cell level **(Fig. 2).** Langerhans cells (LC), being antigen-presenting cells, recognize their own auto-antigen and present it to Th cells. These later become activated as a result of the interaction and start produce an non-typical for the normal immune response spectrum of cytokines and surface receptors. In their turn, LC also become activated and produce together with Th cells the cytokine that activate endothelial cells (EC) of vessels, which become contracted. This increases intercellular spaces and attracts other immune systems cells to the inflammation site, including Th cells. Besides, the EC cells start expressing cell adhesion receptors at their surface which also attracts Th cells from the blood stream. The activated Th cells also cause induction of keratinocytes which then start producing various cytokines that affect LC, Th and EC cells as well as the keratinocytes themselves and start expressing adhesion receptors for Th cells. Besides, activated keratinocytes can themselves present auto-antigens to Th cells. Recently, the functional significance of Tc cells with proposed effector role has been discussed. Thus, a self-sustaining mode, with positive feedbacks, is established.

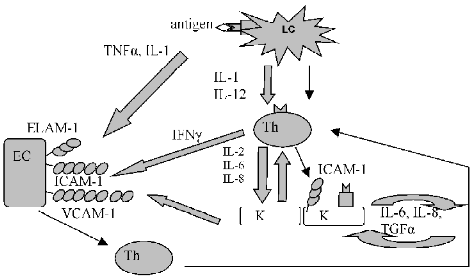


Fig. 2. The model of immune system's pathology (psoriasis). The cell level. K – keratinocytes, EC – endothelial cells, ELAM-1, ICAM-1, VCAM-1 – T-cell adhesion receptors.

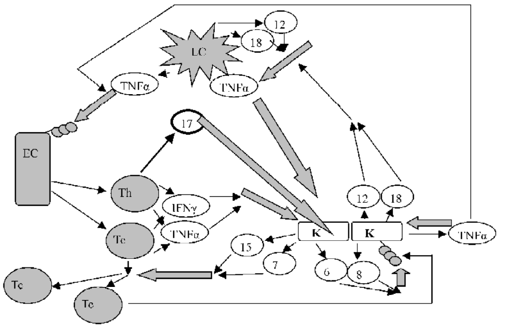


Fig. 3. The model of immune system's pathology (psoriasis). The cytokines level. K – keratinocytes, EC – endothelial cells.

The regulator molecule level **(Fig. 3).** Activated LC produce TNFα, and also IL-12 and IL-18 which, by the positive feedback mechanism, induce TNFα. Th and Tc produce TNFα and IFNγ which activate keratinocytes. Activated keratinocytes, in turn, start expressing a whole spectrum of cytokines. IL-6, IL-8 and TNFα exert effects on keratinocytes themselves. Cytokines IL-12 and IL-18 affect LC, while IL-7 and IL-15 cause proliferation of Tc cells, and IL-7 produced by Th cells induces proliferation of keratinocytes. This is how a self-sustainable mode with positive feedback is established at the level of regulatory molecules.

Results of this study form a basis for computer modeling of a pathological condition. Besides, an adequate presentation of the model needs a systemic description of at least such processes as apoptosis and the cell cycle of keratinocytes.

We are presently studying apoptosis and the cell cycle of keratinocytes in different groups of psoriasis patients. Putting together our own experimental results with data from the literature will, hopefully, allow a more adequate and relatively complete description of such a complex multi-factor systemic disease as psoriasis for the purposes of further computer modeling.

References

Kolchanov N.A. *et al.* Gene Networks // Mol. Biol. (Mosk.). 2000. V. 34. P. 449–460.

EBV INFECTION AND EBV TRANSFORMATION: RECONSTRUCTION OF GENE NETWORKS IN THE GeneNet SYSTEM AND SEARCHING FOR REGULATORY POINTS

Ananko E.A.*, Nedosekina E.A., Oshchepkov D.Yu., Lokhova I.V., Likhoshvai V.A.

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

Keywords: *evolution, genotype, phenotype, mathematical model, computer analysis*

Summary

Motivation: Epstein-Barr virus (EBV) is a common herpes virus that establishes life long persistence in the human host and can directly transform B-lymphocytes. EBV causes the infectious mononucleosis and associates with several human malignancies such as B cell lymphoma, gastric adenocarcinoma, nasopharyngeal carcinoma, and post transplant lymphoproliferative disease. EBV can drive B cell development and survival in the absence of normal B cell receptor signals. The goal of our work is to define the mechanisms by which EBV regulates B cell fate.

Results: Two sections of the GeneNet database were created. The section EBV infection contains description of processes in B cells infected with EBV. The section EBV transformation compiles the information on signal transduction pathways and regulatory proteins as well as on the genes and proteins whose expression changes in transformed B cells. Also, the new section of the TRRD database contains the information on transcriptional regulation of 47 genes involved in processes of EBV infection and/or transformation. Analysis of the information from these new sections revealed several transcription factors important for regulation of B cell fate by EBV. The samples of corresponding binding sites were constructed, and the methods for recognizing binding sites of these factors with help of SITECON were developed. Regulatory regions of genes expressed in B cells were scanned by the methods and some putative binding sites were recognized.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/sections1.shtml

Introduction

Epstein-Barr virus (EBV) causes the infectious mononucleosis and associates with several human malignancies such as Burkitt's lymphoma, Hodgkin's disease, AIDS-associated B-cell lymphoma, primary CNS non-Hodgkin's lymphoma, gastric adenocarcinoma, X-linked lymphoproliferative syndrome, nasopharyngeal carcinoma and post transplant lymphoproliferative disease. *In vitro* EBV transforms B-cells into lymphoblastoid cell lines. Several EBV-encoded proteins let the virus-infected cells to avoid apoptosis (Cohen, 1999).

To clarify the mechanisms by which EBV regulates B cell fate different approaches are used. First, changes in cell phenotype by EBV may indicate what signal transduction pathways will be affected. This could include cell surface receptors and components of signal transduction pathways. Second, to identify some of the cellular proteins that could be targets for EBV-encoded proteins. The first protein, which can interact with EBV-encoded EBNA proteins, was identified as the e-subunit of the human chaperonin TCP-1 complex (Kashuba *et al*., 1999). The second protein was showed to bind EBNA-3 turned out to be the minor subunit of arylhydrocarbon receptor complex (Kashuba *et al*., 2000).

An experimental data obtained using different approaches were collected in the new sections of TRRD (Kolchanov *et al*., 2002) and GeneNet (Ananko *et al*., 2002) databases. An analysis of this information revealed several regulatory points and helps us to develop methods for recognizing regulatory targets in B cells that are affected by EBV.

Methods

For formalized description of interaction of two genomes, EBV and human (B cells), in hybrid gene networks the GeneNet technology (Ananko *et al.*, 2004) was applied.

For creating the samples of natural binding sites for AP-1, ATF2, IRF, Pu.1, NF-κB, NF-Y, Sp1, STAT1 and some other transcription factors in eukaryotic genes the TRRD database (Kolchanov *et al.*, 2002) was used.

For recognition of transcription factor binding sites the SITECON method (Oshchepkov *et al.*, 2004) was employed.

Results

Two new sections of the GeneNet database containing descriptions of hybrid gene networks of interaction of two genomes—EBV and human (B cells)—were created: (1) **EBV infection** contains description of processes in B cells infected with Epstein-Barr virus; (2) **EBV transformation** compiles the information on signal transduction pathways and regulatory proteins as well as on the genes and proteins whose expression changes in transformed B cells. Information content of these two GeneNet sections is given in the Table.

Table. GeneNet database, sections EBV infection and EBV transformation

Class of components	EBV infection	EBV transformation
Compartment	7	4
Process	1	2
Protein	19	75
Reaction	26	102
Gene	9	29
Cell	9	40
Bibliography	35	117
Species	2	2

Three signal transduction pathways playing an essential role in EBV-transformed B cells were described in these sections of the GeneNet database, namely: (1) Jak/STAT signal transduction pathway (activation of the transcription factors Stat1, Stat3, and IRF-7); (2) activation of NF-κB factor by viral protein LMP1 via TRADD/TRAF2; and (3) activation of the MEKK/ERK signal transduction pathway via TRADD/TRAF2. Some other transcription factors are also activated in EBV-transformed B cells, namely AP-1, ATF2, Pu.1, NF-Y, and Sp1. The viral protein EBNA-2 influences transcription of human genes, such as *c-fgr*, *CD21*, *CD23*, *CXCR4*, *CCR6*, and *A20* (data not shown).

For all the listed and some other transcription factors, samples of the corresponding binding sites (natural sites in regulatory regions of eukaryotic genes) were constructed. Regulatory regions of genes expressed in B cells were scanned and some putative binding sites were recognized (data not shown). The full genome of EBV (GenBank entry NC_001345, 172281 bp) was also scanned by the methods. Preliminary results of the recognition for 12 transcription factors are given in Fig. Much more EBV-encoded genes can be regulated by lymphoid/myeloid-specific transcription factors, such as IKAROS, Oct1, Pu.1, than by strictly inducible transcription factors IRF-1, ISGF-3, STAT1, NF-κB (Fig).

Preliminary analysis of the regulatory regions of 47 genes related to EBV infection and EBV transformation annotated in the TRRD database was performed. It was found that the regulatory regions of these genes contain experimentally confirmed binding sites of 89 different transcription factors. The majority of the genes are regulated by transcription factors NF-κB and Sp1 (data not shown).

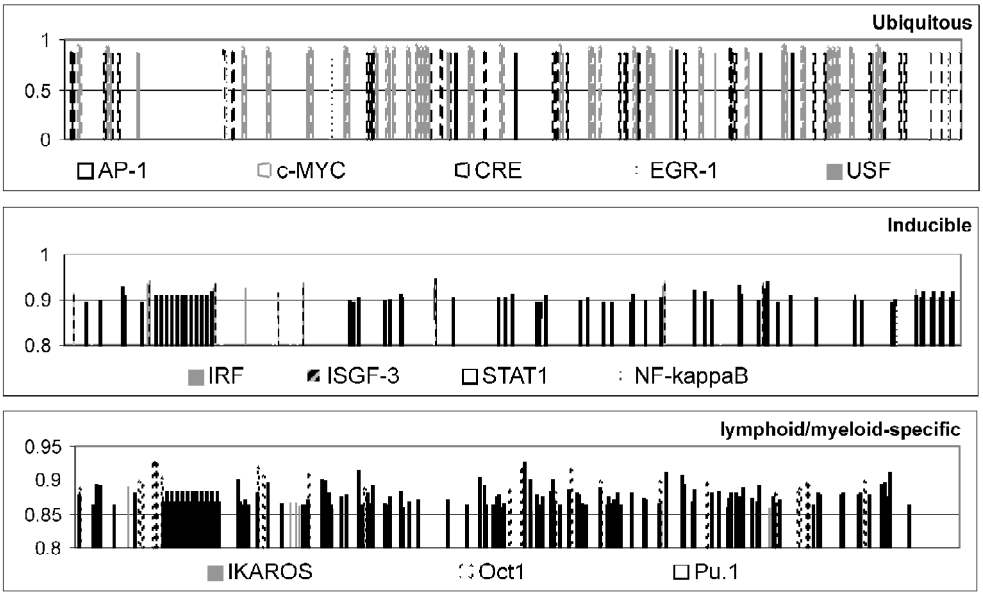


Fig. Predicted binding site for 12 transcription factors in complete genome of EBV (GenBank entry NC_001345, 172281 bp). The binding sites were recognized by the SITECON method (Oshchepkov *et al.*, 2004).

Discussion

As is evident from the preliminary analysis of the information accumulated in GeneNet and TRRD database, activation of signal transduction pathways caused by viral proteins plays a very important role in transformation of B cells by Epstein-Barr virus. In addition, the virus-encoded transactivators, such as EBNA-1 and EBNA-2, capable of stimulating expression of genes of the host cell, are essential for the process in question. Correspondingly, development of methods for recognizing binding sites for transcription factors activated in EBV-transformed B cells and for searching for potential binding sites of these factors in the genes expressed in B cells is the highest priority task. In addition, search for potential binding sites of EBV-encoded transactivators (EBNA-1, EBNA-2, etc.) in regulatory regions of human genes is of great interest.

At present, we are developing methods for recognition of binding sites of viral transactivators EBNA-1, EBNA-2, and some transcription factors, which play an important role in functioning of B-cells (AML-1, AML-2, BSAP, IRF-4, STAT6). Also we began the development of dynamic model of B cell transformation by EBV. Currently, the model contains description of more than 100 dynamic variables.

Acknowledgements

Work was supported by in part by the INTAS Project No. 21-2382, RFBR (03-07-90181-в, 03-04-48469-a, 02-07-90359, 03-04-48506-a), Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology", Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), SB RAS (Integration Projects No. 119), Russian Federal Research Development Program (contract No. 38/2004), NATO (grant No. LST.CLG.979816).

References

Ananko E.A., Loktev K.A., Podkolodny N.L. Development and analysis of models of genetic and metabolic networks and signal transduction pathways in the GeneNet system // Bioinformatics of Genome

Regulation and Structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston, 2004. P. 265–272.

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. GeneNet: a database on structure and functional organisation of gene networks // Nucleic Acids Res. 2002. V. 30. P. 398–401.

Cohen J.I. The biology of Epstein-Barr virus: lessons learned from the virus and the host // Curr. Opin. Immunol. 1999. V. 11. P. 365–370.

Kashuba E., Pokrovskaja K., Klein G., Szekely L. Epstein-Barr virus-encoded nuclear protein EBNA-3 interacts with the epsilon-subunit of the T-complex protein 1 chaperonin complex // J. Hum. Virol. 1999. V. 2. P. 33–37.

Kashuba E., Kashuba V., Pokrovskaja K., Klein G., Szekely L. Epstein-Barr virus encoded nuclear protein EBNA-3 binds XAP-2, a protein associated with Hepatitis B virus X antigen // Oncogene. 2000. V. 19. P. 1801–1806.

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

Oshchepkov D.Yu., Turnaev I.I., Pozdnyakov M.A., Milanesi L., Vityaev E.E., Kolchanov N.A. SITECON – a tool for analysis of dna physicochemical and conformational properties: E2F/DP transcription factor binding site analysis and recognition // Bioinformatics of Genome Regulation and Structure / Eds. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston, 2004. P. 93–102.

FROM GRADIENTS TO STRIPES: A LOGICAL ANALYSIS OF THE GENETIC NETWORK CONTROLLING EARLY DROSOPHILA SEGMENTATION

Chaouiya C.[1]***, Sanchez L.***[2]***, Thieffry D.****[1]

[1] Laboratoire de Génétique et Physiologie du Développement, Marseille, France; [2] Centro de Investigaciones Biológicas, Madrid, Spain
* Corresponding author: e-mail: {chaouiya,thieffry}@ibdm.univ-mrs.fr, lsanchez@cib.csic.es

Keywords: *evolution, genotype, phenotype, mathematical model, computer analysis*

Summary

Motivation: Our aim is to delineate, model and simulate the genetic network controlling the onset of the segmentation process during *Drosophila melanogaster* development.

Results: We present a qualitative modelling approach and its computational implementation (GIN-sim software). Our logical analysis leads to the delineation of the most crucial interactions and regulatory circuits involved in the crucial differentiation decisions at the basis of the segmentation process.

Availability: GIN-sim software is available at the url http://gin.univ-mrs.fr/GINsim.

Introduction

The early embryogenesis of *Drosophila melanogaster* is one of the most extensively studied developmental processes in higher organisms. Saturated mutagenesis followed by careful screening of mutant phenotypes has led to the identification of the key regulatory genes controlling the formation of segments along the anterior-posterior axis of the embryo, prefiguring the specific arrangement of body structures, first in the larva and later in the adult fly [3, 6]. The setting of segmentation involves dozens of genes, expressed either maternally during oogenesis or in the zygotic syncytium. These genes form a hierarchical genetic network, with different modules each responsible for a step in the processing of the initial gradients of maternal products, ultimately leading to specific and robust stripes of zygotic gene expression.

At the onset of the first zygotic gene expression, three maternal products, Bcd, Hb and Cad are each gradually distributed along the anterior-posterior axis. Together, these maternally products define different functional inputs on the gap genes *gt*, *Kr*, *hb* and *kni*. The gap genes are consequently differentially activated along the trunk of the embryo. Furthermore, cross-regulations (predominantly cross-inhibitions) among gap genes amplify these initial differences, ultimately leading to well-differentiated expression domains. At a later stage, the maternal and gap products act together on the pair-rule genes, leading to a further refinement of the segmented gene expression pattern. Finally, all these genes will define the expression of the segment polarity genes, which durably encode the delineation of the segmental borders.

Model

We are progressively modelling the different cross-regulatory modules involved in the control of Drosophila segmentation, using the generalized logical formalism initially developed by R. Thomas and collaborators in Brussels [9, 10, 1, 2]. The logical approach has been implemented in the form of a Java software suite, including a model editor, a core simulator, as well as various analytical tools (graph analysis and layout algorithms, stable states identification, etc.). Using our logical approach and Java software, we have systematically analysed variants of the logical models for the three cross-regulatory modules of the segmentation network, representing the wild type situations as well as various types of perturbations.

Results and Discussion

On the basis of available genetic and molecular data, it proved possible to derive sets of parameters giving rise to simulation results consistent with wild-type as well as with various types of perturbations, including single or multiple loss-off-function mutations, *cis*-regulatory mutations, and ectopic gene expressions [4, 5, 7, 8]. Though many of the corresponding phenotypes have been already experimentally produced, others still await experimental confirmation and thus constitute *bona fide* predictions of the model. Furthermore, our logical analysis outlines the core regulatory components (genes, interactions and regulatory circuits) involved in the crucial differentiation decisions at the basis of the segmentation process.

Acknowledgements

This work was supported by grant PB98-0466 to L. Sanchez from DGICYT, *Ministerio de Ciencia y Tecnologia* of Spain. D. Thieffry acknowledges the financial support from the *Action inter-EPST Bioinformatique*, France. Both group are further supported by a CNRS-CSIC bilateral grant.

References

1. Chaouiya C., Remy E., Mossé B., Thieffry D. // Lecte Notes Control Inf. Sci. 2003. V. 294. P. 119–126.
2. Remy E., Mosse B., Chaouiya C., Thieffry D. // Bioinformatics. 2003. V. 10. ii172–78.
3. Rivera-Pomar R., Jackle H. // Trends Genet. 1996. V. 12. P. 478–483.
4. Sánchez L., Thieffry D. // J. Theor. Biol. 2003. V. 224. P. 517–537.
5. Sánchez L., Thieffry D. // J. Theor. Biol. 2001. V. 211. P. 115–141.
6. St. Johnston D., Nüsslein-Volhard C. // Cell. 1992. V.68. P. 201–220.
7. Thieffry D., Sánchez L. // Ann. NY Acad. Sci. 2002. V. 981. P. 135–153.
8. Thieffry D., Sánchez L. // Curr. Op. Genet. Dev. 2003. V. 13. P. 326–330.
9. Thomas R. // J. Theor. Biol. 1991. V. 153. P. 1–23.
10. Thomas R., Thieffry D., Kaufman M. // Bul. Math. Biol. 1995. V. 57. P. 247–276.

SIGNAL THEORY – AN ALTERNATIVE PERSPECTIVE OF PATTERN SIMILARITY SEARCH

Deyneko I.V.[*1,2], *Kel A.E.*[3], *Wingender E.*[3], *Gössling F.*[2], *Blöcker H.*[2], *Kauer G.*[2,4]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Department of Genome Analysis, GBF, Braunschweig, Germany; [3] BIOBASE GmbH, Wolfenbuettel, Germany; [4] FH Oldenburg/Ostfriesland/Wilhelmshaven, Germany

* Corresponding author: e-mail: blonde@bionet.nsc.ru

Keywords: *signal theory, sequence analysis, pattern detection*

Summary

Motivation: High throughput sequencing techniques involve new strategies and methodologies in both software and hardware design. It is believed that the amount of sequence data doubles every six months, while the number of circuits in computer processors doubles every 18 months. Even if the optimized hardware was at its best, most software utilizes old fashioned character-based algorithms and is not flexible with regard to other kinds of information hidden behind the primary structure of DNA.

Results: We present the implementation of the signal theory based approach for detection of pattern similarity on a genomic scale as described in (Kauer, Blöcker, 2003), investigations of sensitivity and selectivity properties of the method and advise extensions to other kinds of analysis of biomolecules in which signals can be derived from primary biological entities. The entire procedure was accelerated by specific hardware for fast Fourier transformations.

Availability: An experimental internet service for motif recognition will be available (open for public use from May, 2004) at http://genome.gbf.de/

Introduction

To develop entirely new approaches for the analysis of information, encoded in the biomolecules like DNA or proteins, we decided to employ the proven methodologies in image analysis and speech recognition, which deal with information hidden on a very sophisticated level. Clearly, DNA is not present in the cell just as a sequence of four mnemonic letters, but in some spatial structure with lots of chemical and physical interactions and characteristics like charge, hydrophobicity, melting enthalpy etc. Thus, signal theory with its wide variety of methods seems to be a good tool for sequence analysis.

We have developed a novel fast method for detecting sequences on the genomic scale that are similar to a given pattern.

Method

First, let us briefly describe the main steps of the analyses (for detailed theoretical background see (Kauer, Blöcker, 2003, http://genome.gbf.de/wavepaper/). The first step is transformation of the primary biological sequences (example sequence and target sequence (whole genome) into a signal. As a biologically relevant values we used melting enthalpies, which describes melting properties of the double stranded DNA. We also tried artificial coding and suggest other coding schemes as described in the discussion. On the next step we used a rather simple mathematical filter method to compare signals – convolution (Press *et al.*, 1998). The main idea is to calculate a correlation integral (1) and compare it with autocorrelation value (2) for all relevant $\boldsymbol{y}$. Notably, that (1) can be rewritten using Fourier transformants $\boldsymbol{F}$ and $\boldsymbol{G}$ of original functions $\boldsymbol{f}$ and $\boldsymbol{g}$ yielding (3).

$$Corr(y) = \int f(x) \cdot g(x-y)dx \qquad (1)$$

$$AutoCorr = \int g(x) \cdot g(x)dx \qquad (2)$$

$$Corr(y) = InverseFourierT\{F(y) \cdot \overline{G(y)}\}. \qquad (3)$$

Assuming to have direct and inverse Fourier transformations implemented (hardware implementation is best), the entire procedure is reduced just to a multiplication, and scanning for a desired range of values within some predefined threshold.

Implementation and Results

We applied this method for detecting pattern similarity in the following way. As input we take an example sequence (due to specific hardware the length of the pattern is currently limited to 10^6 bp), a target sequence (no length limitations) and a threshold. Then both sequences are transformed into a signal using melting enthalpy values, and the convolution filter is applied. A resulting signal is scanned for matches using the given threshold.

Sensitivity and selectivity. We investigated some properties of this method. Notably, the most crucial step, which mostly defines the accuracy of the entire procedure, is the encoding scheme (used to transform DNA into signal), rather than the fine threshold (to reduce the number of output hits). To show this, we tested both melting enthalpy values and its normalized associates (enthalpy values minus average enthalpy, values given in (Breslauer *et al.*, 1996). One of two identical sequences of 60 bp length each were mutated in a number of positions to reach the predefined value of difference between intact and mutated one (Table). It is easy to see that the normalized values of melting enthalpy are much more sensitive to mutations, since such small changes in the sequence as the presence of just 1 or 2 mutations drastically change the relative correlation coefficient whereas the non-normalized enthalpy appeared to be more tolerant to the mutations. So, the normalized enthalpy is better for more accurate detection of similar patterns, while the non-normalized enthalpy values are better for minimizing false negatives (number of sequences not recognized).

Table. The number of single nucleotide substitutions between the example sequence of a length 60bp and an mutated copy of it

Relative correlation decline	Enthalpy encoded signal	Normalized enthalpy encoded signal
1%	8	1
2%	19	2
4%	40	3
10%	-	6

Another remarkable property of the method is its ability to detect multiple motifs interspersed with a random sequence. We tested our method on composite patterns of the form 50bp-25N-50bp, as well on motifs consisting of three and four conserved parts. In each case the method showed to be robust to the changes of the consecutive order of the conserved parts (for example ABC or ACB or CAB). This feature allows us to use this method for detection of sparse regulatory units like enhancers or promoters.

The performance of the entire procedure was highly accelerated by the special PowerFFT card from Eonic (www.eonic.com) for fast Fourier transformations. All runs were done on a 1,7GHz AMD processor PC with 1 Gb RAM. Figure 1 shows the results for the whole human genome scans for fragments of 10 Kb and 100 Kb. In both cases (shorter and longer pattern), it takes 4 minutes with hardware acceleration to scan through the entire human genome, while the BLAST (character based comparison) speed significantly reduces with the increase of the search pattern length.

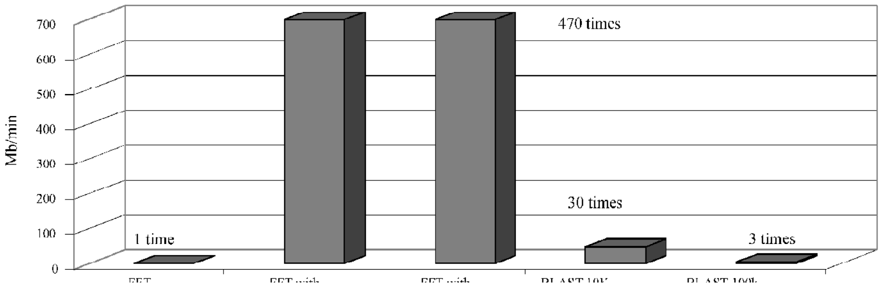


Fig. 1. Total throughput of different search procedures in megabases per minute.

This high throughput platform allowed us to open an worldwide internet service for analyzing sequences based on signal correlation methodology described above. At present, several preloaded genomes (human, mouse, chimpanzee, rat, *E. coli*) are provided, which can be searched using the two described encoding schemes or user defined scheme. External free public access to the experimental server, is scheduled for May 2004.

Discussion

Signal analysis as a different kind of view on DNA. Which of the methods, letter or signal comparison, can provide more prediction accuracy or a wider range of detectable features? It would be completely irrelevant here to test one method using results of another. On the one hand, only wet lab verification or data are the highest authority, on the other hand the space of features is wide enough not to be covered by any of the approaches alone. So the described method is not to be used instead, but in addition to the variety of pre-existing techniques. An excellent example is given in Figure 2. A 60 bp example sequence was scanned through a target random sequence.

```
1 ——GCCCCAGAATAACGACGTC-CGGTGTCCTGTTTTCGTGGACGA-CCGCCTACGTTAACCCAG
2 CCGGCATGTTAGCGCAGACCCGACACGGGCGCCAGTGCCCCTGGGGGGTCCGACTGCGTT——-
            **   *    ** * ***   ** **   * ***  *  *** ** ****

1 GCCCCAGAATAACGACGTCCGGTGTCCTGTTT-TCGTGGACGACCGCCTACGTTAACCCAG
3 GCCGGAAGCTCTCG-TAACCGGTCAGTTGTGCGTCGACCACGATTGCTTTATCAGTGCCAG
  ***  *   *  **    *****    ***   ***   ****  ** *        ****
```

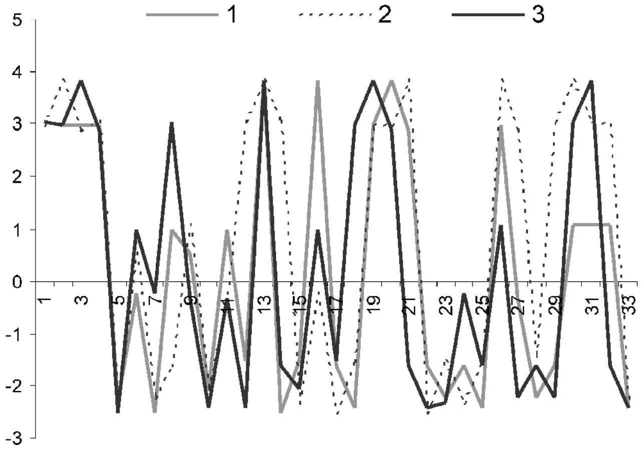


Fig. 2. Sequences in their letter coded and signal representations.

The signal distance between the original pattern (1) and the two detected ones (2 and 3) was very small (less then $3*10^{-6}$, which easy to observe from signals shapes), while as letters they share in common just a few positions (Fig. 2, above – results of pairwise alignment).

It is proposed here, that the signal theory methodology can easily be applied to the analysis of other information bearing macromolecules (for example proteins) or time series of expression data. Suitable filters like low(high)-pass will discover long (short) range regularities in the data. For promoter/enhancer analysis it appears promising to use scores of weight matrices for transcription factors (stored in the TRANSFAC database, ww.biobase.de) to produce a multidimensional signal, which will reveal the similarity in the sense of presence/absence of a specific transcription factor binding site. So the technique presented here may serve as a complement to the existing methods and will evolve in parallel to the latter.

References

Breslauer K.J., Frank R., Blöcker H., Marky L.A. Predicting DNA duplex stability from the base sequence // Proc. Natl Acad. Sci. USA. 1996. V. 83. P. 3746–3750.

Kauer G., Blöcker H. Applying signal theory to the analysis of biomolecules // Bioinformatics. 2003. V. 19. P. 2016–2021.

Press W.H., Flannery B.P., Teukolsky S.A., Vetterling W.T. The art of scientific computing. Cambridge university Press, Cambridge. 1998.

DETERMINATION OF STATIONARY SOLUTIONS IN GENE NETWORK MODELS BY HOMOTOPY METHOD

Fadeev S.I.*[1], Gainova I.A.[1], Berezin A.Yu.[1], Ratushny A.V.[2], Matushkin Yu.G.[2], Likhoshvai V.A.[2,3]

[1] Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [3] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia

* Corresponding author: e-mail: fadeev@math.nsc.ru

Keywords: *genetic systems, modeling, differential autonomous systems, the homotopy method, stationary solutions*

Resume

Motivation: Regarding gene networks as an object of computing and mathematical research we should note that its specific character expresses in very-large-scale systems. The foundation of these systems consists of biochemical processes and phenomena of passive and active transfer of substances and energy between the compartments. Therefore they can be represented as autonomous systems. It generates a need of development of theoretical and numerical methods for analyzing such systems.

Results: In this article the homotopy method is being developed for the construction of stationary solutions of dynamic models of gene networks.

Inrtoduction

The foundation of gene networks consists of biochemical processes and processes of passive and active transfer of substances and energy. If the sphere of gene networks functioning can be split up into some regions in which we can suppose to have a momentary interfusion (homogeneity of space), the modeling can be carried out in terms of autonomous systems of ordinary differential equations of form (Likhoshvai *et al.*, 2003)

$$t > 0, \quad dx_i / dt = F_i(x) - x_i G_i(x), \quad i = 1,2,...,n. \qquad (1)$$

Where x – vector argument with components $x_1,\ldots,x_n$; $F_i(x)$ and $G_i(x)$ – rational functions that take on a non-negative value in case of non-negative values of arguments and parameters of the model.

Characteristics of stationary solutions are known to play an important role in research of autonomous systems. In a given case these characteristics are defined from the following system

$$F_i(x) - x_i G_i(x) = 0, i = \overline{1,n}. \qquad (2)$$

Solutions of (2) must be non-negative with regard to physical meaning. Integration of the autonomous systems (1) is one of the methods of finding solutions of (2). Assuming that t is sufficiently large, the solution of (2) is supposed to become independent of t and gives an approximate solution of nonlinear system (2). Usage of Newton's method for specifying of approximate solution of system (2) raises the reliability of result. Application of this frequently used method is turned to be problematic when the components of autonomous system fall into fast and slow components.

Another possibility of finding stationary solutions is regarded in this article and consists in use of the homotopy method (Oden, 1976; Kholodniok, 1991; Fadeev, 1998). At that the choice of homotopy operator takes count of the structure of equations (2).

Results

We plunge the system (1) into the system of nonlinear equations with parameter λ of the following form:

$$0 \le \lambda \le 1, \quad \lambda \; F_i(x) - x_i G_i(x) \; = \; 0 \, , \; i = \overline{1,n}. \tag{3}$$

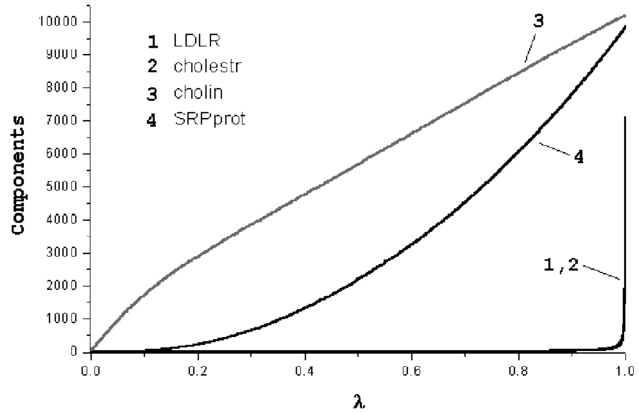


Fig. 1. Parameter continuation λ of system (3) and the rise on stationary solution of autonomous system (1) if λ = 1.

At that when λ = 0 system (3) has a precise solution x = 0. Thus if the solution of (3) exists on a segment $0 \le \lambda \le 1$ we find the solution of system (2) using parameter continuation λ till λ = 1.

In another variant of the homotopy operator assignment, if λ=0, the system of nonlinear equations with parameter λ has a precise solution x^0, where x^0 is an arbitrary vector with components

$$x_i^0, \; i = 1,2,\ldots,n.$$

For that we define vector parameter ε with components ε_i, $i = 1,2,\ldots,n$, using the following conditions:

$$\varepsilon_i \; F_i(x^0) - x_i^0 G_i(x^0) = 0, \qquad i = 1,2,\ldots,n. \tag{4}$$

It is easily seen that the system of nonlinear equations

$$0 \le \lambda \le 1, \quad [\lambda \; + \; (1 \; - \; \lambda)\varepsilon_i] \; F_i(x) - x_i G_i(x) \; = \; 0 \, , \quad i = 1,2,\ldots,n, \tag{5}$$

actually has a precise solution $x = x^0$ if $\lambda = 0$. If λ=1, systems (5) and (2) are congruent.

It is seen from numerical experiments that the construction of parameter dependence of the solution sometimes happens to be difficult when λ falls into the neighborhood of λ=1. In such cases you ought to use the following variants of homotopy operator. Instead of (3) you should use

$$0 \le \lambda \le \infty, \; (1 \; - \; e^{-p\lambda}) F_i(x) - x_i G_i(x) \; = \; 0 \, , \; i = \overline{1,n} \, , \tag{6}$$

where p > 0 is a given parameter;

instead of (5) you take $0 \le \lambda \le \infty$,

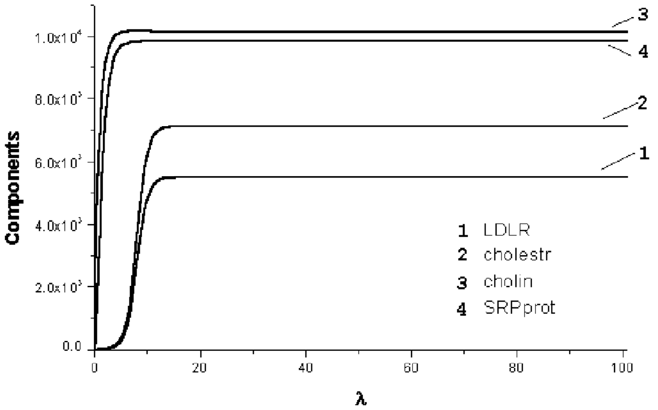


Fig. 2. Parameter continuation λ of the system (6). Neighborhood of λ=1 on the image 1 was "stretched" at the expense of the choice of homotopy operator

$$[1 - e^{-p\lambda} + e^{-q\lambda}\varepsilon_i]\ F_i(x) - x_i G_i(x) = 0\ , i = \overline{1,n}\ , \tag{7}$$

where p>0, q>0 are given parameters. The solution of the system (2) is obviously close to the solution of system (6) or (7) if λ is sufficiently large. We would like to note that the use of homotopy operators (5) or (7) not always allow to find the solution of the system (2) when initial solution was arbitrary chosen. During parameter continuation λ critical point (like "turn") might appear. After passing through this point the extension goes backwards and returns to the zero value of λ. As a case in point we examined a primary model of regulation of the cholesterol intracellular biosynthesis in form of system (1) consisted of 10 equations (Latipov *et al.*, 2003).

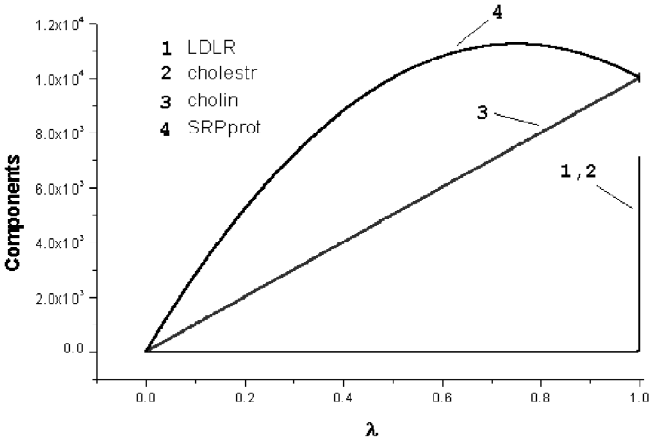


Fig. 3. Parameter continuation λ of the system (5). Initial solution is a vector with unit components.

Applying different variants of homotopy method, we found the solution of the system (2) that turned to be congruent stationary solution received by means of integration of the Cauchy problem. The results are shown on Figures 1–5.

Application of the homotopy method to the system (1) consisted of 39 equations, which represents more complete model of regulation of the cholesterol intracellular biosynthesis (Ratushny *et al.*, 2000), allowed to find one more stationary solution that was proved to be an unstable solution. It is interesting that parameter continuation of system (6) gave the same result as the integration of the Cauchy problem. But the parameter continuation of system (3) led to a new stationary solution.

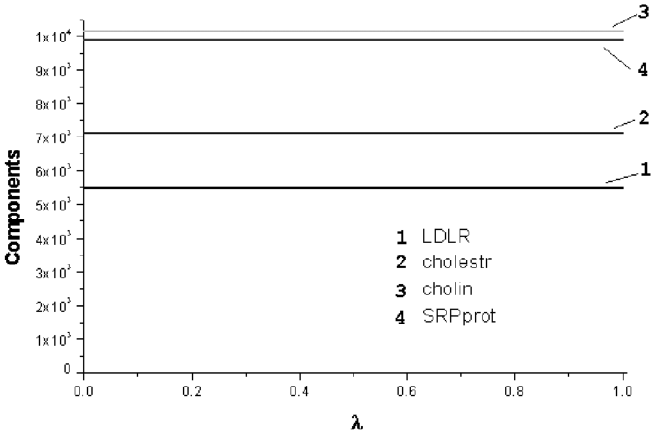


Fig. 4. Parameter continuation λ of the system (5), where initial approximate solution was found by means of the integration of the autonomous system. The strict lines are the proof of the stationarity of the solution.

Using the same example we tested a procedure of diagram construction of stationary solutions. It was made on pattern of the series of the solutions of the system (5) in whish the initial solution was taken as the solution obtained in previous step.

Discussion

During the process of functioning intracellular (concerned as open dynamic systems) necessarily come to some area of phase space where they are stable and where they do not exceed the bounds of the area. Steady state of functioning has either stationary character or oscillatory character. In the first case we observe invariability of concentration of substances and in the second case all the groups of substances undergo essential changes of concentration. Oscillations might have circulating, quasi-periodic and even chaotic character. Nevertheless the chaotic oscillations do not leave the point of attraction. Another particular feature of gene networks is a possibility of existence more that one alternative states of functioning; each of them realizes in proper environment. One of the problems of mathematical modeling of gene networks is a revelation of all the patterns of behavior and conditions of their realization and in particular a search of all non-negative solutions of the concerned system.

As we already mentioned, the integration of the autonomous systems for stationary solution seeking not always might give a reliable result. Thus it is important to have a set of methods for solving this problem. The homotopy method is one of such methods. In this article we offered 4 different homotopy operators. And it is shown that at least one of them leads to rise on stationary solution regardless of its stability. At that right choice of homotopy operator is an essential condition of normal completion of task solution.

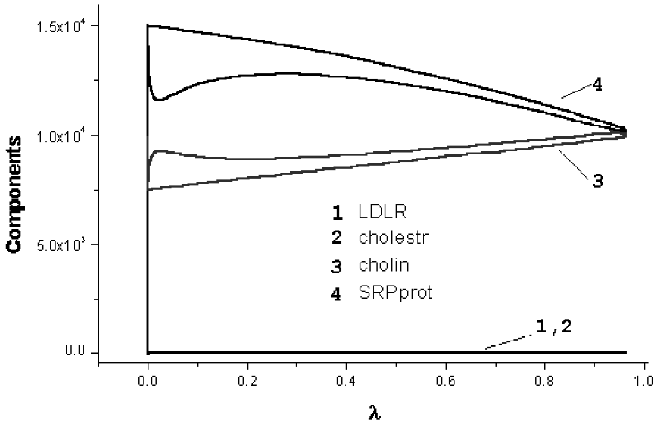


Fig. 5. Example of rise of turning points in the system (5) at using the parameter continuation λ.

We see the further development of this method as follows. First step is a creation of data bank of homotopy operators for definite classes of systems of equations which describe gene networks and also checking of the operators on concrete models. Second step is a development of this method till finding all stationary solutions of the system (2). First of all we would like to develop the methods of ruse on the solution of the system (2) starting from a given initial point and also the methods of definition of initial solutions area from which we can reach the solutions of the system (2). The third step is a development of the parameter continuation method. Such methods allow to define a belonging of finding stationary solutions to the appropriate diagrams in order to have a global idea about the properties of the stationary solutions in combination with the stability.

Thus, if we make an adequate construction of gene-net and make a happy choice of homotopy operator, we can find and analyze such stationary states of real gene networks. But such states extremely rarely become apparent during the experiment (life) because of either rare need in them or fleetingness of the process owing to engaging other mechanisms (for instance, during the process of ontogenesis) or the obscurity of the state for the observer.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (No. 02-04-48802, 02-07-90359, 03-01-00328, 03-04-48506, 03-04-48829, 03-07-96833, 04-01-00458), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4) and "Origin and evolution of biosphere" (No. 10002-251/ п -25/ 155-270/200404-082); SB RAS (Integration Project No. 119).

References

Fadeev S.I., Pokrovskaya S. A., Berezin A.Yu., Gainova I.A. Program package STEP for computational investigation of systems of nonlinear equations and autonomous systems in general form. Novosibirsk: NSU, 1998. 188 p.

Kholodniok M., Klich A., Kubichek M., Marek M. Analyzing methods for nonlinear dynamic models. M.: Mir, 1991. 368 p.

Latypov A.F., Nikulichev Yu.V., Likhoshvai V.A., Ratushny A.V., Matushkin Yu.G., Kolchanov N.A. Problems of Control of Gene Networks in a Space of Stable States // Proc. IFAC Workshop "Modelling and Analysis of Logic Controlled Dynamic Systems". Irkutsk, Russia, 2003. P. 251–266.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. Problems of functioning theory of gene networks // Sib. J. of Industrial Mathematics. 2003. V. 4(14). P. 64–80.

Oden J. Final elements in nonlinear mechanics of continuum. M.: Mir, 1976. 464 p.

Ratushny A.V., Ignatieva E.V., Matushkin Yu.G., Likhoshvai V.A. Mathematical model of cholesterol biosynthesis regulation in the cell // Proc. of the second international conference on bioinformatics or genome regulation and structure. Novosibirsk, 2000. V. 1. P. 199–202.

ABOUT COMPUTATIONAL RESEARCH OF MATHEMATICAL MODELS OF HYPOTHETICAL GENE NETWORKS BY PARAMETER CONTINUATION

Fadeev S.I.* [1] ***, Osokina V.A.*** [2] ***, Likhoshvai V.A.*** [3,4]

[1] Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia; [2] Novosibirsk State University, Novosibirsk, Russia; [3] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [4] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Corresponding author: e-mail: fadeev@math.nsc.ru

Keywords: *genetic systems, modeling, differential autonomous systems, the homotopy method*

Resume

Motivation: Needs in analyzing of gene networks require developments of new analysis methods of mathematical models.

Results: In this article we present the research method of complicated hypothetical gene networks. The main point of this method is to present the initial model as whole of basic models whose properties was studied and to expose of area where the initial model behavior and basic models behavior are phase equivalent.

Introduction

One of approaches, which can be effectively applied for properties study of newly developed mathematical models for gene networks, consists in reduction of a new model to models with known properties. At presence in the gene networks theory sufficiently studied models are models of symmetrical canonical hypothetical gene networks (Likhoshvai *et al*., 2003). It is sufficiently obvious that if we take one of earlier studied symmetrical canonical HGN and change its parameters on a small value then a phase-plane portrait of its behavior won't have any modifications of qualities. Thus we have an idea to use our knowledge about symmetrical canonical HGN properties for researches of properties of more complicated HGN. At the initial stage it is natural to consider constructions not very differing from basic models. And as we fill up the basis of studied HGN we can move to the analysis of more and more complicated constructions. In this article we present analysis method for complicated HGN consisted in reduction of complicated HGN to basic ones. Principe of working we demonstrate on concrete examples of HGN in non-canonical form. We also present a new algorithm of searching stationary solutions of HGN in arbitrarily form which unites the homotopy method and Monte Carlo method.

Results

In this article we consider the problem of numerical investigation of mathematical models for hypothetical gene networks (HGN) which are represented by autonomous equations sets described in article (Likhoshvai *et al*., 2003). One of the ways consists in presentation of considered model as the whole of so called basic models. The properties of the basic models are supposed to be well studied. Then in certain areas of parameter changing properties of considered model will be sufficiently close to the properties of one of the basic models. Analysis method bases on artificial parameter introduction (the homotopy method). As the result we have a possibility of directed study of HGN using parameter continuation where the initial solution belongs to one of basic models. Application of the parameter continuation method allow to study the model parameter dependence of both stationery solutions (nonlinear equations sets) and limit cycles (bounded problems for autonomous systems) regardless of stability character.

In this article we give an example of diagram construction for stationary solutions i.e. construction of parameter dependence of stationary solutions with simultaneous stability determination by means of package STEP algorithms (Fadeev *et al.*, 1998). The parameter continuation method allow to expose bifurcation points on diagram and to determine areas of solutions plurality or areas of auto-oscillations self-excitation (if they exist).

Example 1. Autonomous equations set has the following form:

$$\frac{dx_1}{dt}=\frac{\alpha_1+\alpha_2 x_2^\gamma}{1+\beta(x_3^\gamma+x_2^\gamma)}-x_1,\ \frac{dx_2}{dt}=\frac{\alpha_1+\alpha_2(1-\alpha_3)x_3^\gamma}{1+\beta[x_1^\gamma+(1-\alpha_3)x_3^\gamma]}-x_2,\ \frac{dx_3}{dt}=\frac{\alpha_1+\alpha_2 x_1^\gamma}{1+\beta(x_2^\gamma+x_1^\gamma)}-x_3\,. \qquad (1)$$

Where α_1, α_2, β and γ are nonnegative parameters, $\gamma > 1$. One needs to determine parameters areas where auto-oscillations exist.

For diagram construction of stationary solutions model (1) was "plunged" into the model with artificially entered parameter α_3, $0 \le \alpha_3 \le 1$. This model has next form:

$$\frac{dx_1}{dt}=\frac{\alpha_1+\alpha_2(1-\alpha_3)x_2^\gamma}{1+\beta[x_3^\gamma+(1-\alpha_3)x_2^\gamma]}-x_1,\ \frac{dx_2}{dt}=\frac{\alpha_1+\alpha_2(1-\alpha_3)x_3^\gamma}{1+\beta[x_1^\gamma+(1-\alpha_3)x_3^\gamma]}-x_2,\ \frac{dx_3}{dt}=\frac{\alpha_1+\alpha_2(1-\alpha_3)x_1^\gamma}{1+\beta[x_2^\gamma+(1-\alpha_3)x_1^\gamma]}-x_3 \,.(2)$$

As the result the model research was carried out taking under consideration properties of 2 basic models which are contained in (2):

$M_1(3,3)$ model - $\alpha_2 = 0$, $\alpha_3=0$, and $M_1(3,2)$ model - α_2 0, $\alpha_3=1$.

For stationary solutions diagram construction initial solution was being found by the parameter continuation method relating parameter α_2 from $\alpha_2 = 0$ till $\alpha_2 = 1$ when $\alpha_3=0$. At that we used our knowledge about stationary solutions of $M_1(3,3)$ model (Fadeev *et al.*, 2002a). Then taking into consideration stability of stationary solutions represented on diagrams of the model (2) we discovered the area of changing parameters α_1, α_2 where auto-oscillations exist at all values of parameter α_3, $0 \le \alpha_3 \le 1$. At that we used known $M_1(3,2)$ model properties (Fadeev *et al.*, 2002b). Thus we point out the area of changing parameters α_1, α_2, where typical properties of M_1 (3,2) model were extended to the model (1).

Example 2. Model which is represented by an autonomous system

$$\frac{dx_1}{dt}=\frac{\alpha}{1+\beta[x_3^{\gamma 1}+x_2^{\gamma 2}]}-x_1,\ \frac{dx_2}{dt}=\frac{\alpha}{1+\beta[x_1^{\gamma 1}+x_3^{\gamma 2}]}-x_2,\ \frac{dx_3}{dt}=\frac{\alpha}{1+\beta[x_2^{\gamma 1}+x_1^{\gamma 2}]}-x_3, \qquad (3)$$

is plunged into the model

$$\frac{dx_1}{dt}=\frac{\alpha}{1+\beta[x_3^{\gamma 1}+x_2^{q\gamma 2}]}-x_1,\ \frac{dx_2}{dt}=\frac{\alpha}{1+\beta[x_1^{\gamma 1}+x_3^{q\gamma 2}]}-x_2,\ \frac{dx_3}{dt}=\frac{\alpha}{1+\beta[x_2^{\gamma 1}+x_1^{q\gamma 2}]}-x_3, \qquad (4)$$

which includes artificial parameter q>0. We managed to find the area of changing parameters $\alpha>0$, $\gamma_1> 0$, $\gamma_2>0$, where the model (3) describes auto-oscillations. At that we used known properties of $M_1(3,3)$ model which follows from (4) when $q =\gamma_1/\gamma_2$.

Example 3. We consider model of HGN which has the following form:

$$\frac{dx_1}{dt}=\frac{\alpha}{1+\beta(x_3^\gamma+x_2^\gamma)}-x_1,\ \frac{dx_2}{dt}=\frac{\alpha}{1+\beta x_1^\gamma x_3^\gamma}-x_2,\ \frac{dx_3}{dt}=\frac{\alpha}{1+\beta(x_2^\gamma+x_1^\gamma)}-x_3\,. \qquad (5)$$

In this example we demonstrate new way of determination of all HGN stationary solutions by the parameter continuation relating artificially entered parameter *q* 0<q<1. When q = 0 the initial solution is an arbitrary set of non-trivial numbers x_1^0 , x_2^0 , x_3^0 belonged to the cube with edge which equals α. Toward this end model (5) is plunged into the model containing parameter *q*:

$$\frac{dx_1}{dt}=\frac{\alpha}{1+\beta \quad r_1(x_3^\gamma+x_2^\gamma)}-x_1,\ \frac{dx_2}{dt}=\frac{\alpha}{1+\beta \quad r_2 \quad x_1^\gamma x_3^\gamma}-x_2,\ \frac{dx_3}{dt}=\frac{\alpha}{1+\beta \quad r_3(x_2^\gamma+x_1^\gamma)}-x_3,$$

where $r_i \ = \ q \ + \ (1-q)\varepsilon_i,\ \varepsilon_i=\frac{1}{\beta \ z_i^0}(\frac{\alpha}{x_i^0}-1)$, $i=1,2,3,$

$$z_1^0=(x_3^0)^\gamma+(x_2^0)^\gamma,\ z_2^0=(x_1^0x_3^0)^\gamma,\ z_3^0=(x_2^0)^\gamma+(x_1^0)^\gamma.$$

Numbers x_1^0, x_2^0, x_3^0 were set with aid of the arbitrary numbers sensor. Then the stationary solution of the system (6) is being continued till value q = 1. As result for sufficiently large α and γ we found 3 solutions and only 1 of them is stable.

Conclusion

Development of analysis methods for gene networks is a primary task of biomathematics, area of system biology of postgenom era occupied with mathematical and computer analysis of molecular-gene systems. In this article we present the method of gene networks research by reduction them in some parametric areas to HGN constructions with known properties. A development importance of this direction expresses in the fact that they will find application at solving different practical and theoretical problems, at construction of gene networks theory and gene networks with known in advance properties. Prospects of the further method development we connect most of all with replenishment of studied HGN bank. In the future we plan to investigate the matter of equivalence of different HGN construction functioning using this method. This is very important from the point of view of development of one more significant direction – theory of adequate pressure (reduction) of mathematical models.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (No. 02-04-48802, 02-07-90359, 03-01-00328, 03-04-48506, 03-04-48829, 03-07-96833, 04-01-00458), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), RAS Presidium Programs "Molecular and Cellular Biology" (No. 10.4) and "Origin and evolution of biosphere" (No. 10002-251/ П -25/ 155-270/200404-082); SB RAS (Integration Project No. 119).

References

Fadeev S.I., Klishevich M.A., Likhoshvai V.A. 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). 2002a. V. 2. P. 96–98.

Fadeev S.I., Pokrovskaya S.A., Berezin A.Yu., Gainova I.A. Program package STEP for computational investigation of systems of nonlinear equations and autonomous systems in general form. Novosibirsk: NSU, 1998. 188 p.

Fadeev S.I., Vernikovskaya E.V., Purtov A.V., Likhoshvai V.A. Determination of bifurcational parameter values of mathematical model m(n,k) of hypothetical gene networks // Proc. III Intern. Conf. on Bioinformatics of Genome Regulation and Structure (BGRS'2002). 2002b. V. 2. P. 99–101.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. Problems of functioning theory of gene networks // Sib. J. of Industrial Mathematics. 2003. V. 4(14). P. 64–80.

ON ONE ALGORITHM FOR MODELING PASSIVE TRANSPORT IN CELL SETS OF ARBITRARY CONFIGURATION

Galimzyanov A. V.

Institute of Biology, Ufa Research Center, Russian Academy of Sciences, Ufa, Russia,
e-mail: galim@anrb.ru

Keywords: *control gene networks, passive transport, modeling, linear programming*

Summary

Motivation: When modeling the dynamics of the systems that control the development of Metazoa, one should account for the processes of spatial distribution of different molecular substances participating in the formation and regulation of functional regimes in cellular gene networks, i.e., determining the levels of gene activity in the cells.

Results: On the basis of mathematical tools for linear programming, an algorithm has been elaborated for modeling passive transport of molecular components (RNA, proteins, multimeric complexes) of control gene networks in cell sets (associations) of an arbitrary configuration.

Introduction

Efficient mathematical methods have been developed in recent years for analyzing the dynamics of gene expression control systems (Smolen *et al.*, 2000). The next step in this direction is to simulate the dynamics of cell sets on the basis of cellular gene network models with account for the influence of *intercellular* molecular signals. The present paper is devoted to a novel algorithm for simulating passive transport, which means spatial distribution of molecular substance from domain of higher concentration to domain of lower concentration (according to gradient) without chemical ATP-energy consumption.

Task Setting

Given is a cell set C of cardinality I (number of cells). For the set C we introduce an adjacency function Ψ correlating each element $c_i \in C$ ($i=\overline{1,I}$) with the assembly C_i, whose elements are the c_i-adjacent cells ($C_i \subseteq C$). An adjacent cell is considered to be such that has intercellular interactions (contacts) with the element c_i. In a particular case the function Ψ can be specified by a simple list of the names of adjacent cells for each cell in C. The element $c_i \in C$ is specified with seven symbols:

$$< ID_i,\ \bar{z}_i(t),\ C_i(t),\ M_i,\ \tilde{\mathbf{W}}_i,\ \hat{\mathbf{W}}_i,\ \bar{\xi}_i(t)\ >, \text{ where} \tag{1}$$

ID_i is the cell identificator (name); $\bar{z}_i(t) = [\, z_1^i(t), z_2^i(t), z_3^i(t)\,]$ are Cartesian coordinates of the cell; $C_i(t)$ is the assembly of adjacent cells of cardinality J_i;

$M_i = \{m_1, m_2, \ldots, m_{K_i}\}$ is the list of molecular components (RNA, proteins, multimeric complexes) occurring in the cell; they may be either synthesized within the cell itself or come from the outside. Let us assume it to be similar for all cells, i.e., $M_i \equiv M_j$, $\| M_i \| = K$;

$\tilde{\mathbf{W}}_i$ and $\hat{\mathbf{W}}_i$ are the weight coefficient matrices of dimension K x J_i, where the elements $\tilde{w}^i_{k,j}$ and $\hat{w}^i_{k,j}$ denote the interaction force (degree of intercellular contact) between the cell c_i and its j-th neighbour by the substance m_k in two directions – from the cell c_i to the j-th neighbour ($\widetilde{w}^{i}_{k,j}$) and from the j-th neighbour to the cell c_i ($\hat{w}^{i}_{k,j}$);

$\bar{\xi}_i(t)=[\xi^i_1(t),...,\xi^i_k(t),...,\xi^i_K(t)]$ is the concentration of relevant molecular components in the cell at an instant of time t.

Thus, the cell set Ω is specified with (1) type elements. A combination of sets $C_i(t)$ as well as matrices $\widetilde{\boldsymbol{W}}_i$ and $\hat{\boldsymbol{W}}_i$ ($i=\overline{1,I}$) determine the cellular network configuration. For the time span under consideration let us suppose the following: (a) the number of cells remains the same; (b) spatial coordinates of all cells are fixed (cells do not move); (c) adjacency relations remain unchangeable. The conditions (a-c) imply that the cellular network configuration is stable. The state $\Omega(t)$ of the system C at an instant of time t is described with a set of vectors $\bar{\xi}_i(t)$ ($i=\overline{1,I}$), i.e., the matrix $\Xi(t) = \| \xi^i_k(t) \|$ dimension I x K, whose element $\xi^i_k(t)$ is the substance concentration k in the cell i at an instant of time t. Matrix $\Xi(t)$ represents a pattern of concentration distribution of the molecular components from the assembly M throughout the cells from the assembly C at an instant of time t. Initial conditions for the system C are specified with the matrix $\Xi(t_0)$. Thus, the description of the system C dynamics consists in obtaining a sequence of matrices $\Xi(t_1)$, $\Xi(t_2)$, …, $\Xi(t_j)$, … over a given time span of observations.

Algorithm

Further, we shall consider an arbitrary cell c_i (let it be the *base cell*), a set of its adjacent cells $C_i=\{c_{i_1}, \ldots, c_{i_j}, \ldots, c_{i_{J_i}}\}$ and substance m_k.

Let us select three non-overlapping subsets out of the set:

$C^D_i(t, m_k)=\{c_{i_j} : \xi^{ij}_k(t) > \xi^i_k(t), j\in 1,...,J_i\}$ is the set of cell donors for c_i;

$C^A_i(t, m_k)=\{c_{i_j} : \xi^{ij}_k(t) < \xi^i_k(t), j\in 1,...,J_i\}$ is the set of cell acceptors for c_i;

$C^E_i(t, m_k)=\{c_{i_j} : \xi^{ij}_k(t) = \xi^i_k(t), j\in 1,...,J_i\}$ is the set of “neutral” cells for c_i.

The sets $C^D_i(t, m_k)$, $C^A_i(t, m_k)$, $C^E_i(t, m_k)$ have cardinalities J^D_i, J^A_i, J^E_i, for which the following equality is held: $J^D_i + J^A_i + J^E_i = J_i$. Let us enumerate the cells of $C^A_i(t, m_k)$ from 1 to J^A_i. For better visualization let us denote cells from $C^A_i(t, m_k)$ with symbols ζ_l ($l=\overline{1,J^A_i}$), the base cell c_i with the symbol ζ^*, local concentrations of substance m_k in these cells with variables $\xi_1(t), \ldots, \xi_{J^A_i}(t), \xi^*(t)$, respectively.

Let us state the problem on the molecular distribution of substance between the reference and adjacent cells in terms of the problem on linear programming (LP) with mixed constraints (2):

Maximize the linear form $\Phi(x)=\sum_{l=1}^{J^A_i} x_l$ for a set of vectors $\boldsymbol{x}=(x_1,\ldots, x_{J^A_i})$ satisfying the conditions:

$i)\, x_l \geq 0 \quad (l=\overline{1,J^A_i})$; $ii)\, x_l \leq \widetilde{w}_l[(\xi^* - \sum_{j\neq l} x_j) - \xi_l] \quad (l=\overline{1,J^A_i})$, where

x_l is the movement of substance m_k from the base cell ζ^* to the cell ζ_l at an instant of time t; $\widetilde{w}_l \in [0, ½]$ is the weight coefficient (analogue for diffusion or filtration coefficient) that signifies the interaction contact force between the base cell and the l-th cell acceptor by substance m_k in the direction from ζ^* to ζ_l.

Constraints (i) and the range of weight coefficient values $\widetilde{w}_l$ represent the mass transfer process in the direction of averaged concentrations of the substance in adjacent compartments. According to constraints (ii), the number of molecules moving of a cell donor to some specific adjacent cell acceptor depend on the weight coefficient value by this substance and on the difference between the concentration levels of the substance in two adjacent cells and cannot exceed $(\xi^* - \xi_l)\,\widetilde{w}_l$. Besides, the transport of the substance from ζ^* to ζ_l can only be possible if the difference between its concentration level in the base cell and collective movement of the molecules of this substance to other cell acceptors ($C_i^A(t, m_k) \backslash \zeta_l$) exceeds $\xi_l(t)$ at the current step t.

General scheme of the algorithm

1. Each cell c_i from C in the metabolic profile $\Xi(t)$ is sequentially selected as a base cell (ζ^*). Then we form relevant sets of cell donors $C_i^D(t, m_k)$ and cell acceptors $C_i^A(t,m_k)$.
2. By solving the LP problem of (2) type we calculate the maximum output of gene product from the base cell at the t-th step – number $\varphi_k^i(t) = \max\limits_{(x_1,\ldots,x_{J_i^A})} \Phi(\mathbf{x})$, and also the portions of this product for adjacent cell acceptors – vector $\mathbf{x}^*(t)$.
3. Upon completion of sequential steps (1-2) we have a set of $\psi_k^1, \psi_k^2, \ldots, \psi_k^i, \ldots, \psi_k^I$: $\psi_k^i = \psi_k^{i,1} + \psi_k^{i,2} + \ldots + \psi_k^{i,j} + \ldots + \psi_k^{i,J_i^D}$, where $\psi_k^{i,j}$ is the number of the molecules of substance m_k that move from cell i to cell $j \in C_i^D(t, m_k)$ at the t-th step. Thus, ψ_k^i is the total gain of substance m_k in the cell i as a result of its transfer from all adjacent cell donors at the t-th step.
4. $\Delta_k^i(\mathrm{t}) = \psi_k^i(t) - \varphi_k^i(t)$ is the change in the amount of substance m_k in the cell i as a result of the system operation at the t-th step.
5. We recalculate molecular distribution of the substance among all the elements in the cell set, i.e., a metabolic profile is formed: $\Xi(t+1)$: $\xi_k^i(t+1) = \xi_k^i(t) + \Delta_k^i(t)$ $(i=\overline{1, I})$.

Stop criterion. Steps (1-5) should be executed until the condition is fulfilled: $\psi_k^i(t) = 0$, $\varphi_k^i(t) = 0$ $\forall i$ $(i=\overline{1, I})$.

Implementation and Results

The algorithm is realized as a computer program in the concept of object-oriented programming and appears, as a calculation module, in the computer program package "**A**nalyzer of the **Ge**ne **N**etwork **Dy**namics" (**AGENDY**) intended for *in silico* analysis of the dynamics of gene-molecular systems controlling gene expression (Galimzyanov, 2000). The LP problem is solved with the aid of the simplex method (Dantzig *et al.*, 1955). By way of example (Fig.) we give calculations for the dynamics of a cell set in the unlocked orthogonal configuration with weight coefficients similar to all intercellular contacts ($\widetilde{w}_{k,j}^i = 1/5$, $\hat{w}_{k,j}^i = 1/5$).

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	1000	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	2500	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

$\Xi(t_0)$

37,5	39,0	36,7	32,2	28,2	24,7	23,7	21,0	18,7	15,8	12,7	9,4	5,8	2,3	0,0
38,9	38,6	36,7	31,8	29,7	26,8	25,8	22,9	20,6	16,8	14,1	10,5	6,8	3,2	0,0
35,5	35,8	36,2	34,2	31,5	29,2	27,7	24,9	21,2	17,4	13,8	11,3	7,5	4,2	1,0
31,0	31,9	33,1	35,1	22,7	31,0	28,2	25,9	21,9	18,7	15,9	12,5	8,9	6,0	1,9
27,9	29,3	31,6	32,8	35,1	32,8	28,8	25,1	22,7	19,0	15,6	12,4	10,0	6,5	1,9
24,6	27,2	29,4	30,3	32,9	32,9	29,7	26,0	22,8	18,9	16,0	13,0	9,5	5,8	2,2
22,3	25,0	27,2	28,5	29,9	29,2	29,2	26,1	22,6	19,4	16,0	14,2	10,1	7,2	2,2
20,3	22,8	24,0	24,9	25,5	27,5	26,7	25,5	21,9	18,0	15,5	13,2	9,4	6,2	2,2
18,1	19,6	20,1	21,1	22,6	23,7	23,2	21,9	21,3	18,8	14,5	12,0	8,5	5,4	2,0
14,9	16,8	17,9	17,9	19,0	19,7	19,7	19,2	18,4	17,0	14,6	11,3	8,0	4,8	0,9
12,4	13,6	14,6	14,8	15,8	15,6	15,4	15,8	14,8	13,9	12,6	10,1	6,9	3,6	0,0
9,2	9,9	10,1	11,1	12,6	12,4	12,2	12,9	12,2	11,4	9,2	9,3	5,6	2,4	0,0
5,5	5,5	6,8	7,8	9,4	9,8	9,9	9,9	8,7	7,3	7,0	4,5	4,7	1,1	0,0
2,2	2,3	3.4	4.4	5,1	5,8	6,4	5,8	5,7	4,9	4,2	2,3	1.1	0.0	0.0
0.0	0.0	0.0	1,0	1,0	2.3	2,3	2.3	2.1	0.9	0.0	0.0	0.0	0.0	0.0

$\Xi(t_s)$

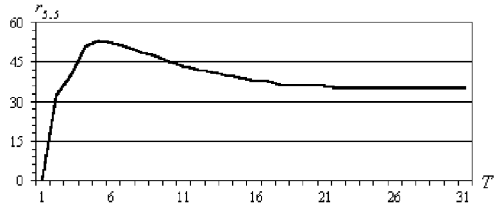


Fig. Dynamics of a hypothetical cell set. $\Xi(t_0)$ is the metabolic profile of the system at an instant of time t_0; $\Xi(t_s)$ is the stationary state; T is the number of iterations (t_s=31); $r_{ij}(t)$ – concentration of the substance (molecules/cells) in the cell c_{ij} at the t-th step. The graph shows concentration dynamics of the substance in the cell $c_{5,5}$ (encircled).

Discussion

The elaborated algorithm falls in the category of machine cellular automata (Toffolli, Margolus, 1987). A comparative evaluation of the efficiency of cellular automata models for diffusion was performed in (Bandman, 1999). The results of the algorithm operation using hypothetical cellular networks under different initial conditions satisfy the following characteristics: constant amount of the substance over the whole period of observations, molecular distribution of the substance among compartments in the direction of spatial-averaged concentration, stationary-state evolution of the system in a finite time span (convergence of the process). Thus, the process under study is quite adequately described with the proposed algorithm. Its integration into the formalism of generalized threshold models (Tchuraev, 1991) will make it possible to simulate the dynamics of control gene networks with account for passive transport within cell sets of an arbitrary configuration.

References

Bandman O.L. Comparative study of cellular-automata diffusion models // Lecture Notes in Computer Science. 1999. V. 1662. P. 395–409.

Dantzig G., Orden A., Wolfe P. Generalized simplex method for minimizing a linear from under linear inequality constraints // Pacific J. Math. 1955. V. 5. P. 183–195.

Galimzyanov A.V. Software automated package for analyzing the dynamics of control gene networks // Proc. of the BGRS'2000. Novosibirsk, 2000. V. 1. P. 233–234.

Smolen P., Baxter D.A., Byrne J.H. Modeling transcriptional control in gene networks: methods, recent results, and future directions // Bull. Math. Biol. 2000. V. 62. P. 247–292.

Tchuraev R.N. 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. 1991. V. 151. P. 71–87.

Toffolli T., Margolus N. Cellular automata machine. MIT Press, Massachussetts. 1987.

CLOSED TRAJECTORIES IN THE GENE NETWORKS

Golubyatnikov V.P.[1]*, *Makarov E.V.*[2]

[1] Institute of Mathematics SB RAS, Novosibirsk, Russia; [2] Siberian Department of International Institute for Nonlinear Science RAS, Novosibirsk, Russia
* Corresponding author: e-mail: glbtn@math.nsc.ru

Keywords: *dynamic systems, Hopf theorem, fixed point theorem, mathematical model*

Summary

Motivation: Multistability is an important property of gene network functioning. Estimating of the possible numbers of limit cycles and stationary points of dynamical systems is a fundamental problem of theoretical and applied mathematics.

Results: We prove the existence of limit cycles for some classes of the gene networks models.

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

Introduction

Detection of closed trajectories in any particular dynamic system is in general a very difficult mathematical problem, even in the low-dimensional cases. Some its particular cases are related to the classical Hilbert's 16-th problem and to the Poincare's Center-Focus problem. Here, we consider special dynamic systems as models of the gene networks. We study their periodic trajectories and stationary points. The existence of these regimes is very important from the viewpoint of the gene networks design meeting the needs of biotechnology, biocomputationg and gene therapy, see (Elowitz, Leibner, 2000; Gardner *et al.*, 2000; Golubyatnikov *et al.*, 2003).

Model

For the models of the gene networks introduced in (Likhoshvai *et al.*, 2001), we consider here corresponding 3-dimensional dynamic systems.

$$\frac{d\ x_i}{d\ t} = \frac{\alpha}{1 + x_{i-1}^{\gamma}} - x_i;\ \alpha>0;\ i = 1,2,3. \qquad (1)$$

$$\frac{d\ x_i}{d\ t} = \frac{\alpha}{1 + x_{i-1}^{\gamma} + x_{i-2}^{\mu}} - x_i\ ;\ \alpha>0;\ i = 1,2,3. \qquad (2)$$

$$\frac{d\ x_i}{d\ t} = \frac{\alpha}{1 + x_{i-1}^{\gamma} \cdot x_{i-2}^{\mu}} - x_i;\ \alpha>0;\ i = 1,2,3. \qquad (3)$$

We assume that $\gamma > \mu > 1$; i-1=3, i-2=2 for i=1, and i-2=3 for i=2. Each dynamic system (1), (2), (3) is symmetric with respect to cyclic permutation of the variables $x_3 \to x_1 \to x_2 \to x_3$. We shall occasionally use notations $x_1 = x$, $x_2 = y$, $x_3 = z$.

Linearizations of these systems near their stationary points are described by the matrix

$$A = \begin{pmatrix} -1 & -p & -q \\ -q & -1 & -p \\ -p & -q & -1 \end{pmatrix} \qquad (4)$$

One of its eigenvalues $\lambda_1 = -1 - p - q$ corresponds to the vector (1,1,1). For $p \neq q$, the other eigenvalues λ_2, λ_3 of A are complicated and 2 Re $\lambda_{2,3} = p + q - 2$. All the trajectories of the systems (1), (2), (3) eventually enter the cube $Q = [0,\alpha] \times [0,\alpha] \times [0,\alpha] \subset R^3$ and do not leave it. The diagonal Δ={x=y=z} contains exactly one stationary point $M_*^{(j)}$ of each of these system. Here the index *j* corresponds to their equation numbers (1), (2) or (3).

Results and Discussions

1. The behavior of the trajectories of the system (1) is much more simple than in (2) and (3). Let *r*(*X*) be the vector, which joins any non-diagonal point *X* with its projection onto Δ and let *v*(*X*) = $\frac{dr(X)}{dt}$. The vector product [*r*(*X*), *v*(*X*)] is parallel to Δ. For any non-diagonal point *X*, the coordinates of [*r*(*X*), *v*(*X*)] are strictly negative. Thus, we obtain

TEOREM 1. *All the trajectories of the system (*1*) outside the diagonal* Δ *turn around* Δ *with positive angular velocity.*

It follows from this theorem that outside the diagonal Δ, the dynamical system (1) has no stationary point. Its linearization in a small neighborhood of its stationary point $M_*^{(1)} \in \Delta$ with the coordinates $x_*^{(1)}$ is described by (4) with $q = \gamma(x_*^{(1)})^{\gamma+1}\alpha^{-1}$, $p = 0$.

For $\alpha(\gamma - 2) < \gamma \;\; x_*^{(1)}$, the real parts of $\lambda_{2,3}$ are negative, in this case the stationary point $M_*^{(1)}$ is the unique attraction of all the trajectories of (1). If $\alpha(\gamma - 2) > \gamma \;\; x_*^{(1)}$, then $M_*^{(1)}$ is not stable. The angular velocities of the trajectories of the system (1) outside of small neighborhood *U*(Δ*)* of Δ are bounded away the zero. For a positive value t_0, each point in *D=Q\U(*Δ*)* makes at least one complete turn around Δ during t_0. Let $T_1 = D \cap H_1\,\{x_1 > x_2 = x_3\}$ $T_2 = D \cap H_2\,\{x_2 > x_1 = x_3\}$ and $T_3 = D \cap H_3\,\{x_3 > x_2 = x_1\}$. According to theorem 1, the trajectory of each point $M \in T_3$ arrives to T_1 at a moment $t_1(M) < t_0$. Let $\tau_1 : T_3 \to T_1$ be the shift in the points of T_3 along the trajectories. At some $t_1(M) + t_2(M) < t_0$, the point *M* arrives at T_2. Let $\tau_2 : T_1 \to T_2$, $\tau_3 : T_2 \to T_3$ be analogous shifts. Later, at the moment $t_1(M) + t_2(M) + t_3(M) < t_0$, the point *M* returns to T_3 for the first time. Let $\varphi_1 : T_1 \to T_3$, $\varphi_2 : T_2 \to T_1$, $\varphi_3 : T_3 \to T_2$ be the rotations of compact contractible sets T_i around Δ by the angle 120^0.

Consider now the composition $\varphi_1 \circ \tau_1 : T_3 \to T_3$ of continuous mappings τ_1 and φ_1. The fixed point theorem implies that there is a point $M_0(x_0, x_0, z_0) \in T_3$ such that $\varphi_1 \circ \tau_1(M_0) = M_0$, or, equivalently, the shift $\tau_1(M_0)$ of this point is obtained by rotation of M_0 around the diagonal Δ. Since the system (1) is symmetric with respect to $x_3 \to x_1 \to x_2 \to x_3$, the composition $\varphi_2 \circ \tau_2 : T_1 \to T_1$ maps the point $\tau_1(M_0)$ to itself, hence, the shift $\tau_2 \circ \tau_1(M_0) = (x_0, z_0, x_0)$ coincides with the rotation $\varphi_2^{-1} \circ \varphi_1^{-1}$ of M_0. Finally, $\varphi_3 \circ \tau_3 : T_2 \to T_2$ maps the point $\tau_2 \circ \tau_1(M_0)$ into itself, hence, $\tau_3 \circ \tau_2 \circ \tau_1(M_0)$ coincides with the result of the complete turn $\varphi_3^{-1} \circ \varphi_2^{-1} \circ \varphi_1^{-1}(M_0)$, and we obtain

THEOREM 2. *For Re $\lambda_{2,3} > 0$, the dynamic system (1) has at least one periodic trajectory symmetric with respect to the permutation of the variables.*

It should be reminded that the fixed point theorem does not ensure the uniqueness and stability of this periodic trajectory.

2. Some results on the uniqueness and stability of such a cycle can be obtained from the Hopf bifurcation theorem, see (Marsden, McCracken, 1976).

In contrast with the dynamic system (1), the behavior of trajectories of the systems (2) and (3) is much more complicated. Trajectory of the system (3) in Figure 1 does not have a constant direction of rotation around Δ. Here $\alpha = 3.237$, $\gamma = 1.725$, $\mu = 1.434$.

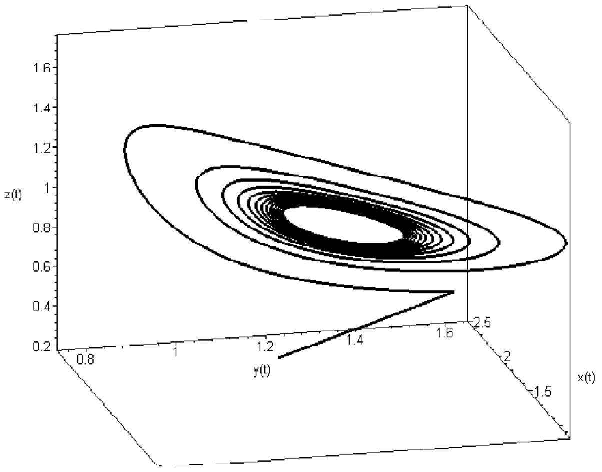


Fig. 1. Limit cycle of the system (3).

The systems (2) and (3) can have three stable stationary points in the neighborhoods of the non-diagonal vertices of the cube Q and three unstable stationary points outside Δ.

Linearization of the system (2) in small neighborhoods of its stationary point $M_*^{(2)}$ is described by the matrix (4) with $p = \mu \cdot (x_*^{(2)})^{\mu+1}\alpha^{-1}$, $q = \gamma \cdot (x_*^{(2)})^{\gamma+1}\alpha^{-1}$. Similar expressions can be derived for the system (3). For each of our systems, if Re $\lambda_{2,3} = 0$ at $\gamma = \gamma_0$, then $\frac{d}{d\gamma}(Re\ \lambda_{2,3}) > 0$ at $\gamma = \gamma_0$. The Hopf bifurcation theorem implies that for $\gamma > \gamma_0$ sufficiently near γ_0, some small neighborhood of the point $M_*^{(i)}$, $i = 1,2,3$, contains a periodic cycle of the system (1), (2) or (3). Figure 2 shows the convergence of two trajectories of the system (2) to the limit cycle from outside and inside. Here $\alpha = 5.908$, $\gamma = 2.981$, $\mu = 2.0$.

Similar phenomena were observed in the system (3). If $\alpha > 2$ and $\gamma + \mu > 4$, this system does not have the Hopf bifurcation. Some of our constructions can be accomplished in R^n, $n > 3$.

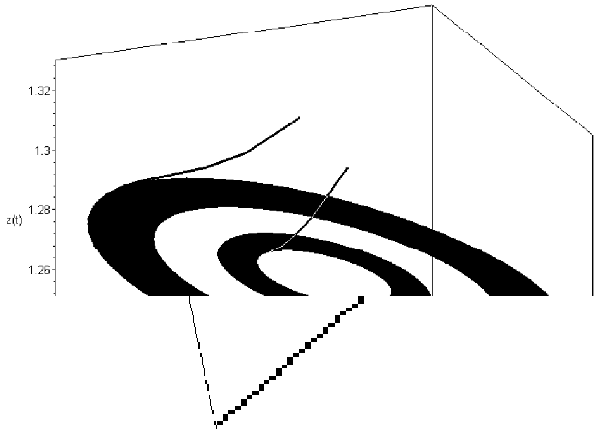


Fig. 2. The Hopf bifurcation in the system (2).

Acknowledgements

The work was supported by the leading scientific schools grant N 311.2003.1 of the President of the Russian Federation, by the grant No. 03-01-00328 of the RFBR and by the interdisciplinary grant No 119 of SB RAS. The authors are indebted to J.P. Francoise and Y.N. Yomdin, and especially to V.A. Likhoshvai and K.V. Storozhuk for helpful discussions.

References

Elowitz M., Leibner S. A synthetic oscillatory kinetic method for modeling gene network // Nature. 2000. V. 403. P. 335–339.

Gardner T., Cantor C.R., Collins J.J. Construction of a genetic toggle switch in *Escherichia coli* // Nature. 2000. V. 403. P. 339–342.

Golubyatnikov V.P., Likhoshvai V.A., Fadeev S.I., Matushkin Yu.G., Ratushny A.V., Kolchanov N.A. Mathematical and computer modeling of genetic networks // Proc. International Conference Human &Computers. 2003, University of Aizu, Japan, 2004. P. 200–205.

Likhoshvai V.A., Matushkin V.A., Fadeev S.I. Relationship between a Gene Network Graph and qualitative Model of its Functioning // Mol. Biol. 2001. V. 35, N 6. P. 926–932.

Marsden J.E., McCracken M. The Hopf bifurcation and its applications. Springer-Verlag, Berlin, 1976.

HOPF BIFURCATION IN THE GENE NETWORK MODELS

Golubyatnikov V.P.[1]*, Volokitin E.P.[1], Osipov A.F.[2], Likhoshvai V.A.[3,4]

[1] Institute of Mathematics SB RAS, Novosibirsk, Russia; [2] Novosibirsk State University, Novosibirsk, Russia; [3] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [4] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Corresponding author: e-mail: glbtn@math.nsc.ru

Keywords: *dynamic systems, Hopf theorem, fixed point theorem, mathematical model, negative feedback, regulation of mRNA degradation, regulation of protein degradation*

Summary

Motivation: Multistability is an important property of gene network functioning. Estimation of the possible numbers of limit cycles and stationary points of gene networks with various types of regulation is a fundamental problem of applied mathematics.

Results: We prove the existence of limit cycles for four classes of the gene networks models where the protein concentration control is at the level of the regulation of their stability. We show that the change in regulation from the stage of the gene expression activation to that of the degradation of the products of the synthesis does not affect the dynamical properties of the gene networks under consideration.

Introduction

Detection of closed trajectories in any dynamic system is a difficult mathematical problem even, in the low-dimensional cases. Here, we consider special dynamical systems as models of the gene networks. We study their periodic trajectories and stationary points. The existence of these regimes is very important from the viewpoint of the gene networks design for the needs of biotechnology, biocomputation and gene therapy, see (Elowitz, Leibner, 2000; Gardner *et al*., 2000; Golubyatnikov *et al*., 2003). We analyze here the gene networks models based on the regulation of the degradation stages of the synthesis products of the genetic elements in contrast with (Golubyatnikov, Makarov, 2004) where regulation is effected at the stages of the initiation of mRNA and protein synthesis. We show that in both cases the qualitative properties of the corresponding dynamic systems are similar and depend on the general structure of the gene network, rather than on the particular work of the regulatory mechanism.

Model

For the models of the gene networks introduced in (Likhoshvai *et al*., 2001), we consider the corresponding 3-dimensional dynamic systems.

$$\frac{d\,x_i}{d\,t} = \alpha - x_i(1 + x_{i-1}^{\gamma}) \;;\ \alpha>0;\ i = 1,2,3. \qquad (1)$$

$$\frac{d\,x_i}{d\,t} = \alpha - x_i(1 + x_{i-1}^{\gamma} + x_{i-2}^{\mu})\,;\ \alpha>0;\ i = 1,2,3. \qquad (2)$$

$$\frac{d\,x_i}{d\,t} = \alpha - x_i(1 + x_{i-1}^{\gamma} \cdot x_{i-2}^{\mu})\;;\ \alpha>0;\ i = 1,2,3. \qquad (3)$$

$$\frac{d\,x_i}{d\,t} = \alpha - x_i(1 + x_{i-1}^{\gamma}) \cdot (1 + x_{i-2}^{\mu})\;;\ \alpha>0;\ i = 1,2,3. \qquad (4)$$

We assume that $\gamma > \mu > 1$; *i*-1=3, *i*-2=2 for *i*=1, and *i*-2=3 for *i*=2. Each dynamic system (1), (2), (3), (4) is symmetric with respect to cyclic permutation of the variables $x_3 \rightarrow x_1 \rightarrow x_2 \rightarrow x_3$. We shall refer in the present paper to the equations from (Golubyatnikov, Makarov, 2004) in the following way: [1], [2] etc. Usually, the similar equation numbers will correspond to similar dynamic systems. Note, that the negative feedback in equations (1) – (4) is effected by the degradation terms while in the equations [1] – [3] the inhibition processes are effected by the terms, which are related to the synthesis of the proteins.

Results and Discussions

1. Linearization of each of these systems near its stationary points is described by the matrix [4]. One of its eigenvalues $\lambda_1 = -1 - p - q$ corresponds to the vector (1,1,1). For $p \neq q$, the other eigenvalues λ_2, λ_3 of *A* are complecated and 2Re $\lambda_{2,3} = p + q - 2$. All trajectories of the systems (1), (2), (3), (4) eventually enter the cube $Q = [0,\alpha] \times [0,\alpha] \times [0,\alpha] \subset R^3$ and do not leave it. The diagonal $\Delta = \{ x_1 = x_2 = x_3 \}$ contains exactly one stationary point $M_*^{(j)}$ of each of these system. Here the index *j* corresponds to their equation numbers (1), (2), (3), (4).

2. The behavior of the trajectories of the system (1) is studied exactly in the same way, as in [1] in (Golubyatnikov, Makarov 2004). Similar calculations of the vector products show that
All the trajectories of the system (1) outside the diagonal Δ *turn around* Δ *with positive angular velocity.*
Hence, outside the diagonal Δ, the dynamic system (1) has no stationary point, so, using the same arguments concerning the fixed point theorem, we see that
For Re > 0, the dynamic system (1) has at least one periodic trajectory symmetric with respect to the permutation of the variables.
Note, that the fixed point theorem does not ensure the uniqueness and stability of this periodic trajectory.

3. Certain results on the uniqueness and stability of such a cycle can be obtained from the Hopf bifurcation theorem. For the systems (1) and [1], the condition Re $\lambda_{2,3} = 0$ is equivalent to $\alpha = \frac{\gamma}{\gamma - 2} \cdot \left(\frac{2}{\gamma - 2} \right)^{1/\gamma}$. Fix any $\gamma > 2$. Direct calculations show that $\frac{d}{d\alpha}(\mathrm{Re}\,\lambda_{2,3}) > 0$ for both systems (1) and [1]. More advanced analysis shows that the Lyapunov parameter v_1 is negative here. Hence, the Hopf bifurcation theorem, see (Kuznetzov, 1995), implies the uniqueness and stability of the cycles, which appear in the systems (1), [1] at their bifurcation points.

We see that for any fixed γ and μ, if Re $\lambda_{2,3} = 0$ at $\alpha = \alpha_0$, then $\frac{d}{d\alpha}(\mathrm{Re}\,\lambda_{2,3}) > 0$ at $\alpha = \alpha_0$ for each of all the systems (2), (3) and (4) as well. As above, it follows from the Hopf bifurcation theorem, that for $\alpha > \alpha_0$ sufficiently close to , some small neighborhood of each stationary point $M_*^{(i)}$, *i* =2, 3, 4 contains a periodic cycle of the corresponding dynamical system (2), (3) or (4). The Figure demonstrates the convergence of two trajectories of system (2) to the limit Hopf cycle from outside and inside. The coordinates of the starting points of these trajectories are $x = 1.285$, $y = 1.285$, $z = 1.29$ for the exterior trajectory and $x = 1.27$, $y = 1.27$, $z = 1.31$ for another , and α=5.908, γ=2.981, μ=2.0.

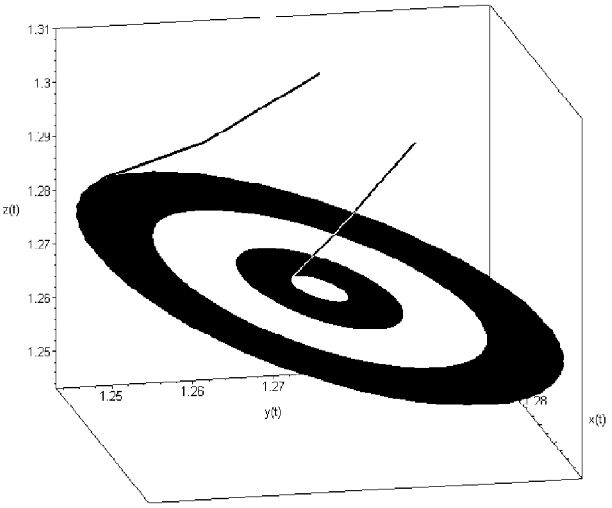


Fig. The Hopf bifurcation in system (2).

Similar phenomena were observed in systems (3) and (4). If $\alpha>2$ and $\gamma+\mu>4$, system (3) does not have the Hopf bifurcation, exactly as the system [3]. More detailed analysis of the trajectories in this Figure shows that their angular and the linear velocities are higher than in the case of system [2]. Analogous arguments show that the Hopf bifurcation can be accomplished in the 5-dimensional dynamic systems composed by formula (1) – (4) and [1] – [3] for $i = 1,\dots,5$. In these cases, the Hopf cycles appear on the 2-dimensional central manifolds corresponding to the eigenvalues of matrix [4] with the maximum real parts.

Acknowledgements

The work was supported by the leading scientific schools grant No. 311.2003.1 of the President of Russia, by the grants No. 01-019-00194, 02-07-90359, 02-04-48802, 03-01-00328, 03-04-48506, 04-01-00458 of the RFBR, by the interdisciplinary grant No. 119 and the project No. 10.4 of SB RAS. The authors are indebted to J.P. Fran3oise, Y.N. Yomdin, and K.V. Storozhuk for helpful discussions, and to A.V. Ratushny for the technical assistance.

References

Elowitz M., Leibner S. A synthetic oscillatory kinetic method for modeling gene network // Nature. 2000. V. 403. P. 335–339.

Gardner T., Cantor C.R., Collins J.J. Construction of a genetic toggle switch in *Escherichia coli* // Nature. 2000. V. 403. P. 339–342.

Golubyatnikov V.P., Likhoshvai V.A., Fadeev S.I., Matushkin Yu.G., Ratushny A.V., Kolchanov N.A. Mathematical and computer modeling of genetic networks // Proc. International Conference Human&Computers-2003. University of Aizu, Japan, 2003. P. 200–205.

Golubyatnikov V.P., Makarov E.V. Closed trajectories in the gene networks // 4-th International conference BGRS-2004. Abstracts. Novosibirsk, 2004.

Kuznetzov Yu.A. Elements of Applied Bifurcation Theory. Springer-Verlag, New-York, 1995.

Likhoshvai V.A., Matushkin V.A., Fadeev S.I. Relationship between a Gene Network Graph and qualitative Model of its Functioning // Mol. Biol. 2001. V. 35, N 6. P. 926–932.

A SOFTWARE ARCHITECTURE FOR DEVELOPMENTAL MODELING IN PLANTS: THE COMPUTABLE PLANT PROJECT

Gor V.[1], *Shapiro B.E.**[1], *Jönsson H.*[2], *Heisler M.*[1], *Venugopala Reddy G.*[1], *Meyerowitz E.M.*[1], *Mjolsness E.*[3]

[1] California Institute of Technology, Pasadena, CA, USA; [2]Lund University, Lund, Sweden; [3]University of California, Irvine, CA, USA

* Corresponding author: e-mail: bshapiro@caltech.edu

Keywords: *Arabidopsis, cellerator, developmental modeling, mathematica, meristem, SBML, systems biology*

Summary

Motivation. We present the software architecture of the Computable Plant Project, a multidisciplinary computationally based approach to the study of plant development. *Arabidopsis thaliana* is used as a model organism, and shoot apical meristem development as a model process. Meristems are the plant tissues where regulated cell division and differentiation lead to plant parts such as flowers and leaves. We are using green fluorescent proteins to mark specific cell types and acquire time series of three-dimensional images via laser scanning confocal microscopy. To support this we have developed an interoperable architecture for experiment design that involves automated code generation, computational modeling, and image analysis.

Results. Automated image analysis, model fitting, and code generation allow us to explore alternative hypothesis *in silico* and guide *in vivo* experimental design. These predictions are tested using standard techniques such as mutants and altered hormone gradients. The present paper focuses on the automated code generation architecture.

Availability. http://www.computableplant.org

Introduction

Scientists who probe the functionality of dynamic developmental systems often express their models mathematically; to make precise system-specific predictions these models are typically encoded with high-level computer languages and standard support libraries and solved numerically. However, high-level languages and libraries typically trade efficiency for generality, and thus may not be appropriate for large hybrid dynamical systems. They also typically lack state-of-the-art technologies in such computationally intensive areas as model optimization and fitting. Finally, custom designed systems are rarely interoperable, making it difficult for researchers to disseminate models.

We have developed an architecture aimed at production-scale model inference. We generate simulation code from models specified in biological and/or mathematical language. Other computational tools are used to analyze expression imagery and other data sources, and the simulator combined with nonlinear optimization is used to fit the models to the data. Key elements include: a *mathematical framework* combining transcriptional regulation, signal transduction, and dynamical mechanical models; a *model generation package* (Cellerator) based on a computer algebra representation; *extensions to SBML* (Systems Biology Modeling Language), an exchangeable model representation format, to include dynamic objects and relationships; *a C^{++}code generator* to translate SBML into highly efficient simulation modules; *a simulation engine* including standard numerical solvers and plot capability; *a nonlinear optimize*r; and ad hoc *image processing* and *data mining* tools. This architecture is capable of simulating processes such as intercellular signaling, cell cycling, cell birth and death, dynamic cellular geometry, changing topology of neighborhood relationships, and the interactions of mechanical stresses.

Methods and Algorithms

Models are input in Systems Biology Markup Language (SBML), an XML-based language for exchanging biological models. SBML is currently supported by more than fifty different software packages used by biological modelers and has become the *de facto* standard for exchanging models among the systems biology community (Finney, Hucka, 2003; Hucka *et al.*, 2003). The modeling interface is provided by *Cellerator* (Shapiro *et al.*, 2003), which allows users to specify models in an arrow-based biochem-ical notation, and translates them automatically to differential equations using a variety of different schemes. *Cellerator* produces extended SBML Level 2 code utilizing *MathSBML* (Shapiro *et al.*, 2004). SBML encoded models are parsed into internal data structures with a *libSBML*-based parser (Bornstein *et al.*, 2004).

Several extensions to SBML have been proposed and will likely be adopted in SBML Level 3 (Finney *et al.*, 2004). In particular, SBML Level 2 does not support spatially-dependent models where each biological entity is *individually defined and enumerated*, and further, does not provide any easy way to describe dynamic geometry and variable size models resulting from cell birth, death, and differentiation. Therefore we have adopted (Finney *et al.*, 2003) to describe dynamic topology and connectivity in terms of arrays, and have extended *Cellerator*, *MathSBML* and *libSBML* accordingly.

Implementation and Results

The *automatic code generator* is central to the architecture. It consists of an *inferencer*, a *rule segmenter and optimizer*, and *application code writer* modules (Fig. 1). It queries the parser for SBML structures and produces efficient C++ application code. The resulting C++ code is then compiled into object code optimized for the desired application. The first two modules of the automatic code generator – the *inferencer* and *rule segmenter* – are pre-processors. They are called once for each SBML model, independent of the application software to be generated. The *inferencer* receives parsed SBML structures from the parser and infers element attributes given the element name. This reflects the inverse relationships between SBML elements and their attributes. For example, the extended SBML has a parameter attribute foreach that indicates the compartment; the *inferencer* creates a list of inferred elements, such as the list of parameters in each compartment.

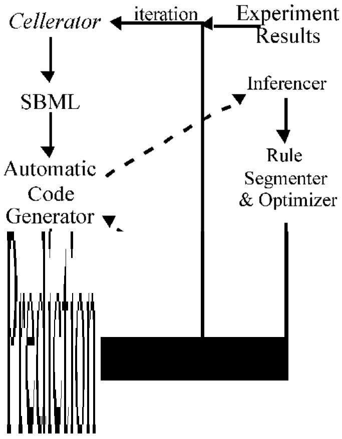


Fig. 1. System architecture.

The *rule segmenter and optimizer* translates SBML rules (which represent mathematical equations using a subset of MATHML) into C++ and performs all necessary renaming of SBML model objects into C variables. Portions of SBML formulas that have no immediate C++ representation, such as the MATHML function sum (which sums a formula over an index) are broken up into sub-rules with intermediate variables; these are later translated into loops or other appropriate control and data structures. Future enhancements will include formula optimization. Identical portions of the formula will be separated into intermediate rules that are only executed once; scalar formulas inside loops will be pre-evaluated outside of the loop. The renaming function completes the work of this module. For example, individual array elements are referenced by index with an SBML model utilizing the MATHML selector operator; this is replaced by the appropriate C array reference such as name[j].

The *application code writer* takes as input the C++ model representation generated by the *rule segmenter and inferencer*, along with an application request, chosen from a menu of available applications. The output is application source code that can be compiled and linked with the chosen application. The *application code writer* consists of a three-level library. The top level contains all of the application-dependent code. This *application level software* is high-level code that is updated as new applications are added. Applications that exist or are being developed include various forward developmental simulators including genetic regulatory network (GRN) temporal synthesis; 4th and 5th order Runge-Kutta differential equation solvers; and optimizers such as Lam-Delosme simulated annealing. In addition, this top level includes overloaded routines that originate at the second level thereby allowing the top level to access this lower level functionality. The second level, *SBML level software,* contains all processes that are not application dependant. This library has entry points for accessing all SBML attributes and elements. The third, and lowest level, is the *utility library*, which contains common operations such as vector algebra and memory maintenance.

Discussion

We are using this simulation environment to extend and enhance our previously reported developmental simulations of the shoot apical meristem (SAM) (Jonsson *et al.*, 2003; Mjolsness *et al.*, 2004]. Our working hypothesis is that SAM development can be described by the differential expression of key regulatory proteins such as CLV1 (a receptor kinase), CLV3 (thought to be the CLV1 ligand), WUS (a transcription factor negatively regulated by CLV1), and a layer-1 specific protein (L1SP). The dependence of CLV1 and CLV3 on WUS, perhaps through a hypothetical diffusible intermediary X, has been inferred from experiments. A second diffusive signal originates from L1SP and diffuses into the rest of the meristem via messenger Y. CLV3 is turned on only if the sum X+Y exceeds threshold. Finally, an unknown diffusible messenger Z creates a surface specific expression pattern for L1SP, which is itself inhibited by STEM, a hypothetical gene expressed only in the lowest meristem layer.

The computable plant architecture provides a systematic, highly automated technique for predictive model generation. The approach combines computer-algebraic represen-tations of biological and mathematical models to produce efficient and problem-specific simulation code. This code can be immediately linked with a menu of external solvers and quantitative predictions generated from the resulting simulations. This architecture is scalable and directly applicable to large-scale developmental systems such as the SAM. The use of extended SBML ensures that models will be interoperable, reusable, and readable by others. Novel to this approach are connections to external solvers by way of automatic code generation and the ability to interpret and solve any biological developmental or cellular process via automatic generation of mathematical and computational tools. Thus no labor is expended writing and debugging problem-specific code, allowing researchers to spend more time on the wet bench.

Acknowledgements

This work was supported by the United States National Science Foundation (NSF) under a Frontiers in Integrative Biological Research (FIBR) grant. HJ was in part supported by the Knut and Alice Wallenberg Foundation through Swegene. Portions of the research described in this paper was performed at the California Institute of Technology.

References

Bornstein B., Keating S., Hucka M., Finney A. libSBML: A Software Toolkit for the Systems Biology Markup Language (SBML) // Poster presentation at Pacific Symp. Biocomp. 2004. (PSB-2004), http://www.sbml.org/libsbml.html

Finney A., Hucka M. Systems Biology Markup Language: Level 2 and Beyond // Biochem. Soc. Trans. 2003. V. 31. P. 1472–1473.

Finney A., Gor V., Bornstein B., Mjolsness E. Systems Biology Markup Language (SBML) Level 3 Proposal: Array Features, 2003. http://www.sbml.org/wiki/arrays.

Finney A., Hucka M., Bornstein B.J., Keating S., Shapiro B.E., Matthews J., Kovitz B., Funahashi A., Schilstra M., Doyle J.C., Kitano H. Evolving a Lingua Franca and Accompanying Software Infrastructure for Computational Systems Biology: The Systems Biology Markup Language (SBML) Project // IEE Systems Biology. 2004. (in press).

Hucka M., Finney A., Sauro H.M., Bolouri H., Doyle J.C., Kitano H., Arkin A.P., Bornstein B.J., Bray D., Cornish-Bowden A., Cuellar A.A., Dronov S., Gilles E.D., Ginkel M., Gor V., Goryanin II., Hedley W.J., Hodgman T.C., Hofmeyr J.H., Hunter P.J., Juty N.S., Kasberger J.L., Kremling A., Kummer U., Le Novere N., Loew L.M., Lucio D., Mendes P., Mjolsness E.D., Nakayama Y., Nelson M.R., Nielsen P.F., Sakurada T., Schaff J.C., Shapiro B.E., Shimizu S., Spence H.D., Stelling J., Takahashi K., Tomita M., Wagner J., Wang J. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models // Bioinformatics. 2003. V. 19. P. 513–523.

Jönsson H., Shapiro B.E., Meyerowitz E.M., Mjolsness E. Signaling in multicellular models of plant development // On growth, form, and computers / Ed. P. Bentley, S. Kumar. Academic Press, 2003.

Mjolsness E., Jönnson H., Shapiro B.E., Meyerowitz E.M. Modeling plant development with gene regulation networks including signaling and cell division // Bioinformatics of Genome Regulation and Structure / Ed. N.A. Kolchanov. Kluwer Publications, 2004.

Shapiro B.E., Hucka M., Finney A., Doyle J. MathSBML: A package for manipulating SBML-based biological models // Bioinformatics. 2004. (In press).

Shapiro B.E., Levchenko A., Wold B.J., Meyerowitz E.M., Mjolsness E.D. Cellerator: Extending a computer algebra system to include biochemical arrows for signal transduction modeling // Bioinformatics. 2003. V. 19. P. 677–678.

A MODEL OF TRYPTOPHAN BIOSYNTHESIS REGULATION

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

Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, 127994, GSP-4, Bolshoi Karetnyi Per., 19, Russia
* Corresponding author: e-mail: gorbunov@iitp.ru

Keywords: *bioinformatics, modeling of biosynthesis, attenuator regulation, repressor regulation*

Summary

Motivation: Modeling of metabolite biosynthesis in bacterial culture and, particularly, constructing a model for gene expression regulation present a long-standing challenge (see Elf *et al.*, 2001; Santillam, Mackey, 2001 for details). Modeling approaches based on differential equation tools or stochastic processes are associated with known computational limitations as well as difficulties in interpreting the result. The latter arise due to indirectness in describing biological processes with such mathematicaltools. Therefore, development of more direct and computationally clear modeling techniques is necessary.

Results: We propose a methodological approach based on the generalized automata theory (directed graphs with potentially infinite number of marks). See the Discussion section too.

Introduction

Let us define the multitude of possible states as a set of vertices of the directed graph with each edge denoting transition between possible states. Each vertex is assigned a rational number (from 0 to 100) corresponding to the percentage of bacteria in culture with the fixed operon (or, more precisely, the orthologous set of operons) being in the state ascribed to this vertex. Phase dynamics on this graph is the change in value distribution of this percentage and average characteristics of the culture.

Let us illustrate the approach with tryptophan synthesis as an example. We consider dynamics of the tryptophan synthesis pathway enzyme concentrations, the metabolite (tryptophan) concentration and the active repressor concentration. To simplify, we consider one enzyme and one repressor, although the model can easily be extended to incorporate several enzymes and repressors/activators. Obviously, metabolite concentrations differ in various bacterial cells at any given instance. Therefore, the model operates with average values (similarly to Elf *et al.*, 2001; Santillam, Mackey, 2001 and other studies) assuming that the average and real concentrations correlate similarly. Concentration heterogeneity can be incorporated into the model by specifying the concentration of each substance at each operon state. The probability of RNA-polymerase binding (at each time phase) is a function of the active repressor concentration; similarly, other probabilities are functions of relevant concentrations. The increase in tryptophan concentration entails (i) the increase in probability of the ribosome shifting from the regulatory sequence, (ii) stronger inhibition of the enzyme by tryptophan, (iii) depletion of the amount of tryptophan synthesized by 1 % enzyme, (iv) buildup of the active repressor. These relationships were expressed so as to incorporate known biological properties and general trends of the biosynthesis. The model also takes into account processes of tryptophan metabolism, decay and transport from the environment, as well as growth of the bacterial culture.

Methods and Algorithms

Our model of tryptophan biosynthesis incorporates (1) a multitude of conditional states of bacteria and the function (perhaps, probabilistic), which, given percentage distribution of bacterial states and current concentrations of the enzyme, tryptophan and active repressor, predicts a new distribution and new substance concentrations for the next phase.

We introduce the terminal state interpreted as “end of game”, which is not assigned the parameter percentage distribution. The amount of bacteria falling into the terminal state is immediately compensated by the income of bacteria at the initial state (“the operon unattended”). Thus, a new percentage distribution is uniquely defined by the old distribution with a set of functions f_q, where for each state q the function $f_q(q')$ determines the percentage distribution of q'-bacteria after transition from state q at the end of the phase. For the sake of simplicity, we let the synthesis success or failure be determined by the succession, in which the ribosome and RNA-polymerase arrive at their corresponding “finishing points”. For ribosome, it is located at a given distance from the regulatory codon area (we assume that the regulatory codons are adjacent; hereafter the regulatory codon is referred to as a codon), and for RNA-polymerase – at a given distance from the pause site. The special case when the ribosome does not bind before the polymerase has reached the “finishing point” we consider a failure. We consider different modifications of function f_q and the function defining concentrations of the enzyme, repressor and tryptophan at a successive phase by those at the preceding phase.

Implementation and Results

We assumed that (1) the procession of RNA-polymerase and ribosome are equal and constant, and (2) the pause time is constant. The time scale (phases) was defined as the time required for processing of one nucleotide by RNA-polymerase. Under this assumption various parameters were taken from the review (Xie *et al*., 2003), private correspondences and roughly estimated from the empirical data. Varying numerical values of ambiguous parameters showed that they have little impact on modeled correlations (at least qualitatively). At the initial phase all bacteria were in state q_0 with zero concentrations of the enzyme, repressor and tryptophan. Over the first 135 phases when the enzyme was not yet produced it had zero concentration, that of tryptophan was close to zero and that of repressor fluctuated around 0.1. Over successive phases the enzyme concentration drastically increased reaching 0.66 in phase 139. In phase 153 the tryptophan content reached 1.68, which is close to its critical level of overexpression. Simultaneously, the repressor concentration also became high 1.75. Over the following phases, all concentrations began to diminish reaching 0.07 for the enzyme and tryptophan at phase 271, which corresponds to underexpression of tryptophan.

Such considerable fluctuations were observed for the initial phases, when the synthesis regulation was yet unstable. During upcoming phases they became smooth and ultimately disappeared. This process was continuous: after 1,500 phases concentrations of the enzyme, tryptophan and the repressor still varied nearly synchronously between 0.08 and 0.3, 0.08 and 0.14, 0.86 and 1.18, respectively. Such fluctuations can be accounted for biologically. Relevant experiments show that ultimate metabolite concentrations do not depend on their initial concentrations and distribution of the cell states in culture.

To compare the attenuator- and repressor-based regulatory strategies, two assays were conducted. In one of them, the contribution of repressor was reinforced (corresponding parameters were set to 0.1, which is two times less than standard values). As a result, fluctuation decay abruptly accelerated: after 1,500 phases they nearly disappeared with the concentration of enzyme stabilizing at 0.11 and that of tryptophan at 0.12. In the other assay, conversely, the repressor contribution to regulation was downgraded (corresponding parameters set to 0.5). This entailed longer and wider fluctuations: after 1,500 phases the concentration of the enzyme and tryptophan varied between 0.06 and 0.14, and between 0.06 and 0.17, respectively. Generally, this suggests that the repressor-mediated regulation is more efficient and responds quicker to expression of tryptophan than does the attenuator-based system. This result is in agreement with the generally accepted view that the repressor-mediated regulation is more sensitive.

In order to build a more realistic model, a certain degree of randomness was introduced in various ways. Thus, to allow for natural variations in cell demand for tryptophan, the corresponding parameter values were varied randomly around its average. This had little impact on modeled correlations, albeit the parameter fluctuations became more stochastic and did not disappear at the limit.

Discussion

The theory of directed graphs with potentially infinite number of marks can be used to generally describe transcription and translation processes and, particularly, processes of repressor- and attenuator-mediated regulation of operon expression and amino-acid biosynthesis. We propose to define a finite number of graph vertices corresponding to "states of the system". Transitions from one state to another are determined by factors like metabolite concentrations and values of certain random variables. For instance, the concentration of tryptophan mediates transition from the "pause-codon" state (when RNA-polymerase is in the last pause phase, and ribosome is at the last regulatory codon) to the "postpause-codon" state (when ribosome remains at the same codon) or to the "postpause-postcodon" state (when ribosome leaves the regulatory codon area). Besides, in each pause phase we dynamically assign graph vertices with numerical values corresponding to the proportion of bacteria in culture that are currently in a given phase along with some averaged characteristics of the culture. This technique was used to *in silico* model tryptophan biosynthesis regulation. The model confirmed, for instance, the empirical observation that the amino-acid concentration triggers a quicker response of the repressor-mediated rather than attenuator regulation system.

Acknowledgements

We are grateful to A.G. Vitreshak and A.V. Seliverstov for helpful discussions.

References

Elf J., Berg O.G., Ehrenberg M. Comparison of repressor and transcriptional attenuator systems for control of amino acid biosynthetic operons // J. Mol. Biol. 2001. V. 313. P. 941–954.

Santillam M., Mackey M.C. Dynamic regulation of the tryptophan operon: A modeling study and comparison with experimental data // Proc. Natl Acad. Sci. USA. 2001. V. 98(4). P. 1364–1369.

Xie G., Keyhani N.O., Bonner C.A., Jensen R.A. Ancient origin of the tryptophan operon and the dynamics of evolutionary change // Microbiol. Mol. Biol. Reviews. 2003. V. 67(3). P. 303–342.

TWO GENE NETWORKS UNDERLYING THE FORMATION OF THE ANTERIOR-POSTERIOR AND DORSO-VENTRAL WING IMAGINAL DISC COMPARTMENT BOUNDARIES IN *DROSOPHILA MELANOGASTER*

Gunbin K.V.*, Omelyanchuk L.V., Ananko E.A.

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

Keywords: *gene networks; triggers; feedback loops; wing imaginal disc compartment boundaries; developmental process.*

Summary

Motivation: Studies on *Drosophila melanogaster* have become most advanced in genetic and molecular terms (The FlyBase Consortium, 2003). Over the past 80 years, abundant molecular biological data explaining the *D. melanogaster* development have accumulated. The situation is favorable for building qualitative and quantitative models for the patterning of the *D. melanogaster* organs.

Results: The "A/P Border" and "D/V Border" sections were generated in the GeneNet database to describe the components of the gene network controlling the development of the wing imaginal disc in *D. melanogaster*. In this work, the information deposited in these sections was analyzed. The regulatory loops providing the development of the wing imaginal disc were considered. It was shown that the regulatory loops can be composed of 3 groups according to function.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/

Introduction

The *D. melanogaster* imaginal discs (IDs) now are the references for studying the molecular-biological processes determining general development (Held, 2002). Furthermore, powerful databases storing the molecular-biological experimental data on the development of the *D. melanogaster* organs have been created (Brody, 1999). To our knowledge, the development of the *D. melanogaster* organs and the whole organism has not been so far modeled in detail (Kyoda, Onami, 1999). The early steps of development modeling include the construction of the GNs describing the molecular-biological processes controlling development, and also a logical analysis of the GNs to identify the constituent blocks crucial for normal development. In this paper, the information stored in the "A/P Border" and "D/V Border" sections of the GeneNet database containing the data on the formation of the anterior-posterior (A/P) and dorso-ventral (D/V) compartments boundaries of the wing ID is analyzed.

Materials and Methods

Two gene networks for the formation of the A/P and D/V wing ID compartment boundaries were reconstructed using the GeneNet technology (Kolchanov *et al.,* 2000). The "A/P Border" and "D/V Border" were reconstructed on the basis of more than 200 articles and contained the information on the genes, mRNAs, proteins (including the transcription factors), external signals, and also regulatory processes. The information was supplied by the references to the literature sources.

Results and Discussion

Reconstruction of the gene networks for compartment boundary formation. The “A/P Border” and “D/V Border” sections of the GeneNet database were developed. In total, the two sections contained 373 objects, 617 relations, more than 80 unique genes, 120 unique protein and protein complexes, and 6 processes.

General features of the gene networks. The cell groups, located along the A/P and D/V wing ID compartment boundaries, express morphogenic proteins coordinating the development of the entire wing. Hence, the cell groups act as the organizers of wing morphogenesis (Held, 2002). In the GNs for the A/P and D/V compartments boundaries formation, the blocks are: 1) the central block containing the selector genes (*en,* for example) and the gene cassettes, whose expression is controlled by these selector genes; 2) the block of the morphogenic genes (HH, WG) expression and their propagation along the imaginal disc; 3) the signaling pathways (N, HH, WG), providing information exchange between cells; 4) the transcription factors (CI, ARM), activated after signaling pathways are turned on, and theirs target gene cassettes controlled by the transcription factors.

Because the described GN blocks for the formation compartment boundaries function, as a rule, at different time and space, the feedback loops have components separated in time and/or space.

The following subsystems are important to provide the function of the GNs for the A/P and D/V compartments boundaries formation: the molecular triggers, the positive feedback loops and the negative feedback loops. Triggers and similar mechanisms turn one functional GNs mode to another, and in so doing provide two types of specific cell differentiation, as well as the morphogene propagation control mechanism. The positive feedback loops are required for maintaining the cells in a differentiated state. The negative feedback loops limit the timing of gene expression during development (Table).

Table. Examples of the important units for the GNs of A/P and D/V wing ID compartment boundaries formation

Block Types	Examples	
Triggers and trigger-like mechanisms	1.	PTC→CI(inh)⇥*ptc*, PTC⇥SMO, HH⇥PTC, SMO→FU→CI(act)→*ptc*→PTC;
	2.	N→SU(H)(inh), N(a)→SU(H)(act)→*vg*→VG→*ct*→CT→*wg*, WG→FZ2→CK→DSH→PKC⇥SGG→ARM→*dl/ser* →SER/DL→N→SU(H)→*vg*→VG→*ct*→CT→***wg;***
Feedback loops with positive interactions	3.	EN→*pc*→PC→*en*;
	4.	EN→*ph*→PH→*en*;
	5.	*wg*→WG→FZ2→CK→DSH→PKC⇥SGG→ARM→*dl/ser* →SER/DL→N(a)→SU(H)(act)→*vg*→VG→*ct*→CT→*wg*;
	6.	*wg*→WG→FZ2→CK→DSH→PKC⇥SGG→ARM→*dl/ser* →SER/DL→N(a)→SU(H)(act)→*ct*→*wg*
Feedback loops with negative interactions	7.	En⇥*en*;
	8.	SER/DL→N(a)→SU(H)(act)→*vg*→VG→*ct*→CT⇥*dl/ser*;
	9.	SER/DL→N(a)→SU(H)(act)→*ct*→CT⇥*dl/ser*;
	10.	SER/DL→N→SU(H)→*en(spl)*→EN(SPL)→*ct*→CT ⇥*dl/ser*;
	11.	AP→*bx*→BX⇥*ap*

Symbol “⇥” refer to inhibition, “→” to activation. The genes are in lower-case italics, proteins are in upper-case; “→*ptc*→PTC”, for example, refer to gene expression activation, “⇥*ptc*” to inhibition.

Triggers and trigger-like mechanisms. The molecular trigger is responsible for the formation of the A/P compartment boundary specific cell type (Fig. A; 1 in the Table). The HH-pathway is activated as a result of the binding of the ligand protein HH to its receptor PTC. As a consequence, the activity of the PTC protein is inactivated, leads to activation of the SMO protein and to change in the functional GN mode (Fig. A). The central block of this trigger is the mechanism by which the transcription factor CI is converted from an activator to an inhibitor. In both forms, the transcription factor CI has the same binding sites in the enhancers of the target genes (for example, *ptc*); hence, the ratio of the concentrations of the CI activator form to the CI inhibitor form determines the cells fate. Another important function of the trigger is to limit the HH morphogene propagation; when PTC protein bound to HH, HH morphogene cannot propagate into the anterior wing compartment.

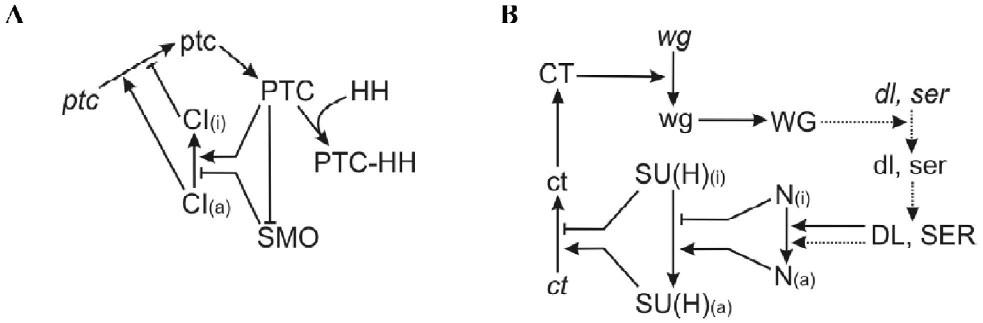


Fig. Molecular trigger controlling the formation of A/P (A) and a similar mechanism controlling the formation of D/V (B) the *Drosophila melanogaster* wing ID compartment boundaries.
Blunt arrowheads refer to inhibition, sharp arrowheads to activation. Discontinuous arrows indicate the processes following those indicated by continuous arrows. The genes are in lower-case italics, RNA is not in italics, proteins are in upper-case.

A trigger-like mechanism control the wing ID D/V compartments boundary specific cell type formation (Fig. B; 2 in the Table). In this system, the central block is also the transcription factor. When the N-pathway is not activated, the transcription factor SU(H) is in the complex together with the H and CTBP proteins and the complex acts as the transcription inhibitor. When the N-pathway is active, SU(H) is bound to the detached intracellular domain of the transmembrane protein N, forming, in such a case, the transcription activator in the complex containing the protein MAM.

Feedback loops, involving positive interactions. To keep the specific A/P boundary cells differentiated, a HH protein inflow is required. The posterior compartment cells have the expression of the selector gene *en* (*engrailed*) maintained from the embryonic stages (3, 4 in the Table). Furthermore, the *en* gene product provides the constant repression of the HH-pathway key components (for example, *ci* and *ptc*), and the activation of the *hh* gene transcription and the transcription of the genes required for the posterior compartment cell type formation.

The *ap* selector gene product, whose expression marks the dorsal compartment specifically suppresses and promotes the expression of genes, thereby making the dorsal compartment cells N protein ligands (SER) different from the ventral (DL). This difference turn on the N-pathway directly in the D/V compartment boundary. The activated N-pathway leads to activation of the transcription of *ct* and *wg*. The WG protein, like HH, is a morphogene, it induces the WG-pathway, this ends up by activating the expression of *dl* and *ser* (5, 6 in the Table) and by inactivating the N protein function. The CT protein is the transcription factor activating the *wg* gene expression and suppressing that of *dl* and *ser* (8–10 in the Table). As a result, the D/V boundary cells express N, and cells at either side of the boundary express SER and DL. Thus, the positive feedback loop is established to maintain the differentiation of the wing imaginal disc D/V boundary cells (5, 6 in the Table).

Feedback loops, involving negative interactions. Although the negative feedback loop cannot overcome obstacles, nevertheless the mechanism controlling the expression of the *en* selector gene operates (7 in the Table). This can contributes both to the autoinhibition of the *en* expression in the A/P compartment boundary and to the elimination of the selector gene expression at a specific time and/or site during the *D. melanogaster* development.

The early stages of the D/V compartment boundary formation are reversible. It is of interest that the *ap* selector gene expression, causing this specific cell type formation, is dynamic. Because the negative feedback loop is free from the inhibition (11 in the Table), the *ap* selector gene expression soon ceases. The *ap* selector gene expression has enough time to start the formation of the specific type of the D/V compartment boundary and to maintain the cells differentiated along this boundary.

Conclusions

In this work, the temporal-spatial integration of the signaling pathways and the morphogene gradients along the D/V and A/P wing imaginal disc compartment boundaries were taken into account in logical analysis. This allowed us to identify the GNs with feedback loops. These loops control the formation of the cell types of the wing morphogenesis organizers located along the wing imaginal disc compartment boundaries (Held, 2002).

Acknowledgements

The work was supported by the RAS Presidium Program “Molecular and Cellular Biology (Project No. 10.4), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501).

References

Brody T. The Interactive Fly: gene networks, development and the Internet // Trends in Genetics. 1999. V. 15. P. 333–334. http://sdb.bio.purdue.edu/fly/aimain/1aahome.htm

Held L.I. Imaginal discs: the genetic and cellular logic of pattern formation. Cambridge: Cambridge University Press, 2002.

Kolchanov N.A., Kolpakov F.A., Podkolodnaya O.A., Ignatieva E.V., Goryachkovskaya T.N., Stepanenko I.L. Gene networks // Russ. J. of Mol. Biol. 2000. V. 34. P. 533–544 (In Russian).

Kyoda K., Onami S. Simulation of genetic interaction for Drosophila leg formation // Pacific Symposium on Biocomputing. 1999. V. 4. P. 77–89.

The FlyBase Consortium. The FlyBase database of the Drosophila genome projects and community literature // Nucleic Acids Res. 2003. V. 31. P. 172–175. http://flybase.org/

COMBINED OPTIMIZATION TECHNIQUE FOR BIOLOGICAL DATA FITTING

Gursky V.V.[2], *Kozlov K.N.**[1], *Samsonov A.M.*[2]

[1] Dept. of computational biology, State Polytechnical University, St.Petersburg, 195251 Russia;
[2] A.F. Ioffe Physico-technical Institute of the Russian Academy of Sciences, St.Petersburg, 194021, Russia
* Corresponding author: e-mail: kozlov@spbcas.ru

Keywords: *optimization, regulatory gene networks, optimal control, mathematical model, computer analysis*

Summary

Motivation: Modern molecular biology has massive amounts of quantitative data already at its disposal. The crucially important problem for getting closer insights into mechanisms of development is to reduce the complexity of finding the parameters of mathematical models by fitting to experimental data.

Results: The new Combined Optimization Technique (COT) showed a high accuracy in reconstruction of phenomenological parameters of equations and saved about 30 % of the most time consuming operations in computation that allow to propose the COT as quite attractive instrument for processing big amounts of experimental data of various nature.

Availability: available on request from the authors

Introduction

Modern molecular biology has massive amounts of quantitative data already at its disposal, and robust and reliable algorithms' development to treat them becomes a foreground job. Mathematical modeling is essential for systematic treatment of experimental results and for getting insights into the structure of underlying natural objects.

We perform the gene expression data fitting in the context of one biological system namely the segment determination gene network of a fruit fly *Drosophila* embryo. The experiments were performed to acquire data on segmentation gene expression at cellular resolution, see (Reinitz, Sharp, 1995). The dynamical model of gene expression is described by a system of highly nonlinear reaction-diffusion (NRD) equations. We present new results of experimental data fitting for finding phenomenological parameters capable of pattern formation and propose a new Combined Optimization Technique (COT) for processing large amounts of experimental data. The developed algorithm combines advantages of random search method and steepest descent approach. Main idea is as follows: firstly a rough approximation of parameters is to be found by the random search, afterwards it is subjected to refinement by the Optimal Steepest Descent Algorithm (OSDA) developed recently (Kozlov, Samsonov, 2003) and applicable to problems of various physical nature.

Main goal is to evaluate the efficiency and convergence speed of the method, and for this reason we start with an estimation of the parameters of small (2 genes) network for simplicity and expanded the number of genes involved after successful numerical experiments. The parameters we have found in 2-gene network allow to keep permanent patterns of gene expression when time tends to infinity. In larger networks the optimization results lead to different asymptotic behavior of solution (Gursky *et al.*, 2004).

Methods and Algorithms

The dynamics of the model is described by the system of coupled differential-difference NRD equations formulated in (Reinitz, Sharp, 1995).

$$\frac{\partial v_n^a}{\partial t} = R^a g\left(\sum_{b=0}^{G-1} T^{ab} v_n{}^b + m^a v_n^{bcd} + h^a\right) + D^a\left(v_{n-1}^a - 2v_n^a + v_{n+1}^a\right) - \lambda^a v_n^a . \quad (1)$$

The equation is written for *each* gene product a and *each* nucleus n, and G is the number of genes. A matrix element T^{ab}, one for each pair of proteins, and coefficients m^a, h^a, R^a, D^a, λ^a for each protein are unknown parameters which should be determined by means of minimization of a functional equal to the sum of squared differences between the concentrations of the gene products (say, proteins), observed experimentally and calculated independently, e.g., by means of gene network approach.

Constraints in the form of inequalities are used to be imposed to the parameters R^a, D^a, λ^a for each protein a, that does not allow to include them directly into an extended Lagrangian. Therefore to apply the Lagrange technique for optimization the constraints are to be transformed into equations, e.g., for R^a as follows:

$$R_{low} \le R^a \le R_{up} \qquad R^a = \alpha_r + \beta_r \tanh(\gamma_r^a r^a) .$$

Constants α and β are defined for parameter R^a by the following formulae

$$\alpha_r = \left(R_{up} + R_{low}\right)/2 \qquad \beta_r = \left(R_{up} - R_{low}\right)/2 .$$

Constraints for other parameters (D^a and λ^a) are transformed similarly, however the transformation is not unique, and other representations involving bounded functions can be used.

Combined Optimization Technique for data fitting consists in application of the Simulated Annealing (SA) method using a weak quality criterion to obtain the rough approximation of the parameter set, which is refined afterwards by the OSDA, see (Kozlov, Samsonov, 2003).

The transformation coefficient values γ_r, γ_d and γ_λ are renewed after each of M steps using an empirical rule. If the functional value has been changed greater than the value of the corresponding parameter during the last M steps, then the corresponding value of γ should be increased in order to make the transformation function (*tanh*) steeper and vice versa.

To make a close comparison of SA and COT we used the quality criterion for COT similar to that which was proposed in (Reinitz, Sharp, 1995) for SA. Namely, when the functional value decreased less than a predefined value θ during the last S steps the set of parameters obtained at the very last step N is identical to the solution of the problem, and the optimal parameters are the components of this vector.

Implementation and Results

The numerical results are given in the Table below. Because of stochastic nature of the SA method the results provided are averaged over all performed experiments. Parameters were recovered with 6.1 % accuracy. The average number of functional evaluations used by COT equals 1.949.198. To obtain similar results with smaller accuracy of 6.9 % using the SA only it required 2.708.485 functional evaluations, that is ~28 % more.

To study the COT convergence in a lab conditions we produced the artificial gene expression data for the network of two genes in eight nuclei by integration the model equations, using the set of parameters that represents already known solution. We took the model output for 9 time moments to calculate the functional value.

We performed optimization for 100 random initial approximations of parameter set $q=\{q_i\}$ and introduced the following criterion κ to measure the precision of numerical simulation q^{opt}:

$$\kappa = \max_i \frac{\left|q_i^{true} - q_i^{opt}\right|}{\left|q_i^{true}\right|} \times 100\% , \qquad (2)$$

where q^{true} is the known solution.

S	κ	Number of functional evaluations
2000	8.9	1932640
1000	8.5	1906387
750	8.7	1908808
500	6.1	1949198
50	9.2	1884532

The weak quality criterion used by the SA part of COT was: $M = 100, S = 5, \theta = 10^{-3}$ and the final one for COT was: $M = 14$, $S = 500$, $\theta = 10^{-9}$. To estimate the efficiency of COT we performed numerical simulations with SA using the following quality criterion: $M = 100, S = 5, \theta = 10^{-5}$.

The average precision of the parameter set obtained by COT is less than 10 % if only those rough approximations produced by SA were taken into account, for which $\kappa < 30$ %. This coincides with the fact that the gradient method converges to the local minimum in general.

Discussion

The widely used Simulated Annealing method converges to the global minimum at the cost of very intensive computations. The number of functional evaluations determines the time necessary to obtain the solution of the data fitting problem because of a huge number of species and, therefore, the differential equations that are to be integrated. In real gene networks this number can exceed three hundreds, and to include more genes in the network under consideration is crucially important for getting closer insights into mechanisms of development.

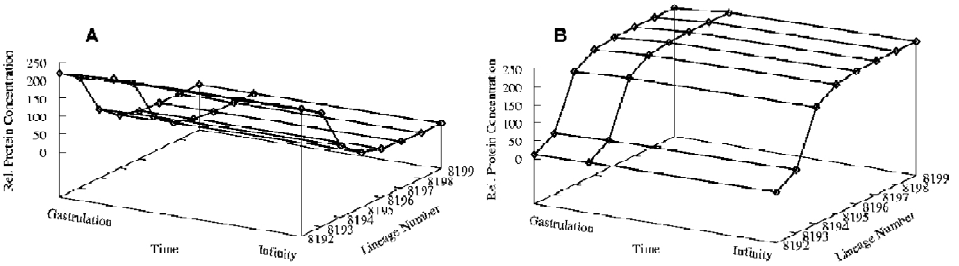


Fig. The time evolution of patterns from gastrulation time to infinity is shown for the first and the second proteins in the network on panels *A* and *B* respectively. The concentration of both proteins in all nuclei does not change in time.

The proposed new Combined Optimization Technique showed a high level of accuracy in reconstruction of phenomenological parameters of equation and saved about 30 % of the most time consuming operations in computation, which may be equal to several days for a large scale problem simulation on a high performance computer. These features allow to propose COT as quite attractive instrument for processing big amounts of experimental data of various nature.

The parameters we have found in 2-gene network allow to keep permanent patterns of gene expression when time tends to infinity. The time evolution of patterns from gastrulation time to infinity is shown in Figure for the first and the second proteins in the network on panels *A* and *B* respectively. In larger networks the optimization results lead to different asymptotic behavior of solution (Gursky *et al.*, 2004).

Acknowledgements

Our sincere thanks to Dr J. Reinitz and J. Jaeger for many valuable discussions. The support of the study by the NIH Grants RR07801, TW01147, the CRDF GAP Awards RBO685, RBO1286 is gratefully acknowledged.

References

Gursky V.V., Jaeger J., Kozlov K.N., Reinitz J., Samsonov A.M. Pattern formation and nuclear divisions are uncoupled in Drosophila segmentation: comparison of spatially discrete and continuous models // Physica D. 2004. (submitted).

Kozlov K.N., Samsonov A.M. New data processing technique based on the optimal control theory // Techn. Physics. 2003. V. 48(11). P. 6–14.

Reinitz J., Sharp D. Mechanism of föormation of eve stripes // Mechanisms of Development. 1995. V. 49. P. 133–158.

MATHEMATICAL MODEL OF THE GENE NETWORK OF TNFα-INDUCED NF-kappaB ACTIVATION

Guryeva Ya.P., Stepanenko I.L., Likhoshvai V.A.*

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

Keywords: *gene networks, mathematical modelling, TNFα, NF-kappaB, apoptosis*

Motivation: Proinflammatory cytokine TNFα (tumor necrosis factor α) induces two important signal pathways: 1) caspase cascade through caspase-8 activation leading to apoptosis, and 2) NF-kappaB transcriptional factor activation, which regulates expression of the genes necessary for cell survival. To investigate impact of these two concurrent processes on each other we developed a mathematical model for TNFα-inducible activation of NF-kappaB.

Results: Using published experimental data and GeneNet technology, gene network of TNFα-inducible activation of NF-kappaB was created. Based on developed network and using generalized chemical kinetic method we constructed a mathematical model which consists of 85 blocks and contains 182 constants and 74 dynamic variables. Using computer analysis of hypothetical mutation effect on gene network function, we obtained quantitative data for the effect of the mutations in cellular-inhibitor of apoptosis proteins (IAPs) and adaptor proteins TRAFs and TRADD on caspase-8 inhibition and activation of NF-kappaB.

Introduction

TNFα is a major mediator of inflammation and nonspecific resistence factor releases on infection event and activates immune system as well as regulates cell proliferation and apoptosis.

The binding of the TNF homotrimer to TNFR1 receptor results in release of the receptor from its complex with SODD inhibitor. This leads to the trimerization of TNFR1 and binding to adaptor protein TRADD. In turn TRADD recruits additional adaptor proteins namely TRAF2 and RIP.

The phosphorylation and degradation of inhibitory protein IkappaB that binds to NF-kappaB in cytoplasm of nonstimulated cell is the requirement for TNF-inducible activation of NF-kappaB. The phosphorylation of IkappaB is carried out by IKK kinase. The activation of this kinase is the result of phosphorylation of IKK by NIK kinase, which takes place after TNF-induced binding of IKK with TRAF2 and RIP proteins. The NIK kinase itself activates by binding to TRAF2.

As a result of IkappaB degradation, free NF-kappaB is transported to nucleus where it binds and regulates expression of wide range of genes. The microarray data shows that during stimulation of TNFα in HeLa cells, more than 100 genes are upregulated with 14 of them regulated by NF-kappaB transcription factor (Zhou *et al.*, 2003).

The adaptor protein TRADD bind to FADD protein on activation of TNF-induced apoptosis pathway. FADD, in turn, recruits procaspase-8, which activates and initiates caspase cascade leading to apoptosis. Activated caspase-8 cleaves many substrates. Some of the latter are proteins TRAF1, RIP and NIK, which are needed for activation of NF-kappaB transcriptional factor.

Although TNFR1 is able to induce caspase-8 activation, TNFα is not cytotoxic for major cell types. This paradox is explained by parallel activation of NF-kappaB pathway, which induces the expression of some antiapoptosis genes including apoptosis inhibitors c-IAP1 and c-IAP2. These inhibitors together with proteins TRAF2 and TRAF1 are suppressing procaspase-8 processing. However activation of caspase-8 and apoptosis pathway occurs on addition of TNFα together with protein synthesis inhibitor CHX. The aim of current work was to develop a mathematical model of TNFα-inducible NF-kappaB activation based on experimental data.

Materials and Methods

For building of mathematical model we used generalized chemical kinetic modelling method (Likhoshvai *et al*., 2000). This method underlies the principle that biological system is conventionally divided on finite number of structural elements (genes, mRNA, proteins and low-molecular compounds) and elementary processes (regulatory interactions and reactions). The elementary processes are described by means of differential equations, which in turn describe the rates of concentration changes of gene network components. By setting up the differential equations for all intercellular processes in a given cell and by applying the law of summation of elementary processes rates, we can unambiguously set up a system of differential equations.

The key stage of model construction is the verification of its parameters. For this purpose the approximate initial values of the parameters are specified, based on literature data analysis. Further on a search for the optimal values for these parameters is performed. At first, all the registered conditions of the individual experiments conducted for the investigated system are reproduced. Then the mathematical model is introduced to these conditions. This introduction is called 'scenario' (Likhoshvai *et al*., 2002). In the end, the values of parameters are selected to reproduce the observed experimental data quantitatively and qualitatively.

Results and Discussion

Using the GeneNet technology we developed a formalized description of gene network of TNFα-inducible activation of NF-κB. The GeneNet database is available through the Internet (http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/). The TNFα-inducible activation of NF-kappaB gene network contains circuits with positive feedback e.g. NF-kappaB induces expression of following genes: a) NF-kappaB subunit p50; b) proteins participating in NF-kappaB activation (e.g. TRAF2); c) proteins inhibiting apoptosis pathway (c-IAP1, c-IAP2, TRAF2 and TRAF1); as well as a circuit with negative feedback: NF-kappaB induces the expression of its own inhibitor gene, IkappaB.

On the base of generalized chemical kinetic method we developed a mathematical model of TNFα-inducible activation of NF-kappaB gene network comprising of 85 blocks and containing 182 constants and 74 dynamic variables. To verify the parameter set of the model the experimental data obtained for HeLa cells were used. The following scenarios were created describing: concentration changes of NF-kappaB in cytoplasm and nucleus under action of TNFα; changes in IkappaB, cIAP1 and cIAP2 concentrations under action of TNFα; and activity change of caspase-8 under action of TNFα and TNFα +CHX.

We compared the results obtained with the mathematical model and experimental data (Fig. 1, 2). As it can be seen from Fig. 1, the translocation of NF-kappaB to nucleus starts in 3–4 minutes after stimulation. The transport is complete in 10–20 minutes after addition of 10 ng/ml TNFα. According to the data presented in diverse experimental studies, approximately 10–20 % of cytoplasmic NF-kappaB is translocated to the nucleus (Ding *et al*., 1998 Verma *et al*., 1995). The result of computation using our model is that 20 % of cytoplasmic NF-kappaB is translocated to the nucleus (Fig. 1).

The caspase-8 is not activated in HeLa cells stimulated by TNFα (Fig. 2). However, activation of caspase-8 occurs in 2 hours after addition of TNFα and protein synthesis inhibitor CHX. At these conditions caspase-8 is able to cut about 80 % of synthetic substrate in 6.5 hours. In presence of CHX the synthesis is blocked for proteins needed for cell survival and coded by genes which expression is regulated by NF-κB.

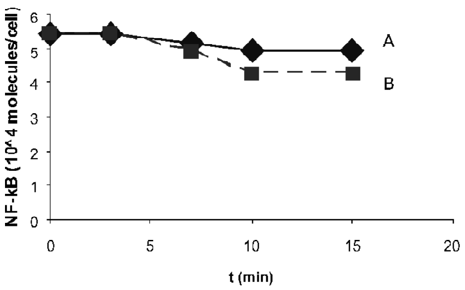


Fig. 1. Concentration change of NF-kappaB in cytoplasm under action of 25 ng/ml TNFα. A– as calculated using mathematical model, B – experimental data (Ding *et al.*, 1998).

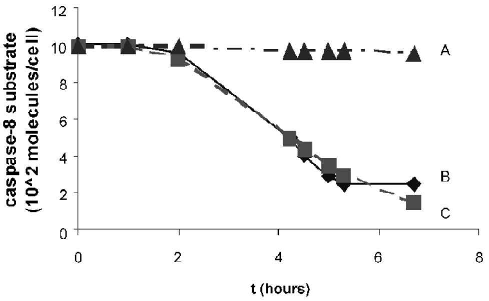


Fig. 2. Caspase-8 activation. A – after addition of 10 ng/ml TNF – as calculated using mathematical model. B – after addition of 10 ng/ml TNFα and CHX – experimental data (Luo *et al.*, 2003). C – after addition of 10 ng/ml TNF and CHX as calculated using mathematical model.

A search of hypothetical mutations that may lead to the increase of caspase-8 activity under action of TNFα was performed. We have simulated the influence of conditional mutations that decrease by one order of magnitude the synthesis constant K_{sin} of proteins cIAP1 and cIAP2 (Fig. 3). In this case caspase-8 cleaves 50 % of synthetic substrate in 6.5 hours. We also investigated the effect of a mutation increasing the dissociation constant K_d of complex TRAF2/TRADD, since FADD, needed for caspase-8 activation, competes with TRAF2 for binding with adaptor protein TRADD. In this case caspase-8 cleaves 30 % of substrate in 6.5 hours.

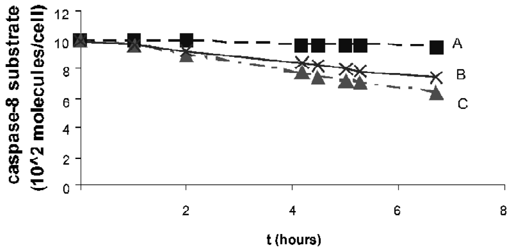


Fig. 3. The influence of model parameters change on caspase-8 activity. The activity of caspase-8 is given after addition of 10 ng/ml TNFα: A – normal conditions; B – with conditional mutation increasing the K_d of TRAF2/TRADD complex (by one order of magnitude); C – in case of conditional mutation decreasing K_{sin} of apoptosis inhibitor proteins cIAP1 and cIAP2.

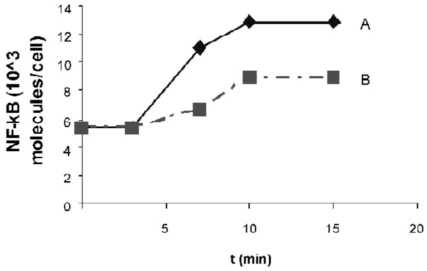


Fig. 4. The influence of hypothetical mutations on NF-kappaB transport. The concentration change of NF-kappaB in the nucleus after addition of 25 ng/ml TNFα. A - normal conditions; B – with conditional mutation, increasing K_d of complex TRAF2/TRADD (by one order of magnitude).

As it can be seen on Figure 3, the mutations that influence such apoptotic proteins as cIAP1, cIAP2 and TRAF2 are leading to the considerable gain of the caspase-8 activity under induction of TNFα. Moreover, in case of increase of K_d for complex TRAF2/TRADD (by one order of magnitude), we observe decrease of NF-kappaB transport to the nucleus (Fig. 4). The amount of active NF-kappaB transcription factor in nucleus is increasing by 2.3 times, whereas in case of mutation this figure reaches only 1.5.

Conclusion

Based on the analysis of published experimental data and using generalized chemical kinetic method of modelling, we developed a mathematical model of TNFα-inducible activation of NF-κB. Using this model one can imitate the functioning of the biosystem in wide range of conditions, and predict the dynamics of gene network behavior. Thus, we have investigated the effect of change in several parameters of the model on fine balance between pathway of caspase-8 activation and NF-kappaB transcriptional factor pathway.

Acknowledgements

The work was supported by the grants No. 03-07-96833, 03-04-48829, 03-04-48506, 03-01-00328, 02-07-90359, 02-04-48802 of the RFBR, by the SB RAS interdisciplinary grants Nos. 119, 148 and the RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4).

References

Ding G.J., Fischer P.A., Boltz R.C., Schmidt J.A., Colaianne J.J., Gough A., Rubin R.A. Miller D.K. Characterization and quantitation of NF-kappaB nuclear translocation induced by interleukin-1 and tumor necrosis factor-alpha. Development and use of a high capacity fluorescence cytometric system // J. Biol. Chem. 1998. V. 273. P. 28897–28905.

Likhoshvai V.A., Matushkin Yu.G., Vatolin Yu.N.., Bazhan S.I. A generalized chemical kinetic method for simulating complex biological systems. A computer model of λ phage ontogenesis // Computational Technologies. 2000. V. 5. P. 87–99.

Likhoshvai V.A., Latypov A.F., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. Technology of using experimental data for verification of models of gene network operation dynamics // Proc. of the Third International Conference on Bioinformatics of Genome Regulation and Structure (BGRS'2002). 2002. V. 2. P. 146–149.

Luo K.Q., Yu V.C., Pu Y., Chang D.C. Measuring dynamics of caspase-8 activation in a single living HeLa cell during TNFalpha-induced apoptosis // Biochem. Biophys. Res. Commun. 2003. V. 304. P. 217–222.

Verma I.M., Stevenson J.K., Schwarz E.M., Van Antwerp D., Miyamoto S. Rel/NF-kappa B/I kappaB family: intimate tales of association and dissociation // Genes Dev. 1995. V. 9. P. 2723–2735.

Zhou A., Scoggin S., Gaynor R.B., Williams N.S. Identification of NF-kappaB-regulated genes induced by TNFalpha utilizing expression profiling and RNA interference // Oncogene. 2003. V. 22. P. 2054–2064.

METABOLIC PATHWAY PREDICTION / ALIGNMENT

Hofestaedt R., Chen M.*

Bioinformatics / Medical Informatics, Technische Fakultaet, Universitaet Bielefeld
Postfach 10 01 31, D-33501 Bielefeld, Germany
* Corresponding author: e-mail: ralf.hofestaedt@uni-bielefeld.de

Keywords: *metabolic pathway, prediction, alignment, PathAligner*

Summary

Motivation: Comparing metabolic pathways are important to identify conservation and variations among different biology system. Alignment is a strong indicator of the biologically significant relationship.

Results: A definition of metabolic pathway is defined. An alignment algorithm and a computational system are developed to reveal the similarities between metabolic pathways.

Availability: http://bibiserv.techfak.uni-bielefeld.de/pathaligner

Introduction

Today a huge amount of molecular data about different organisms has been accumulated and systematically stored in specific databases (Collado-Vides, Hofestaedt, 2002). This rapid accumulation of biological data provides the possibility of studying metabolic pathways systematically. Analysis of metabolic pathways is an essential topic in understanding the relationship between genotype to phenotype (Dandekar *et al.*, 1999).

Researches on genomic sequence alignment have been so far intensively conducted. Applications and tools, such as FASTA [http://www.ebi.ac.uk/fasta3] and BLAST [http://www.ncbi.nlm.nih.gov/BLAST] have been developed to further understand the biological homology and estimate evolutionary distance. Although the information provided by sequenced genomes can yield insights into their evolution and cellular metabolism, knowledge of the genome sequence alone is really only the start point of the real work.

Several approaches of metabolic pathway alignment are already made in the past years. Forst C.V. and Schulten K. (1999; 2001) extended the DNA sequence alignment methods to define distances between metabolic pathways by combining sequence information of involved genes. Dandekar *et al.* (1999) compared glycoslysis, Entner-Doudoroff pathway and pyruvate processing in 17 organisms based on the genomic and metabolic pathway data by aligning specific pathway related enzyme-encoding genes on the genomes. Tohsato Y. *et al.* (2000) proposed a multiple (local) alignment algorithm by utilizing information content that was extended to symbols having a hierarchical structure EC numbers. We are going to present a new pathway alignment strategy to analysis and fully characterize metabolic pathways in the cell.

Metabolic Pathway Definitions

A biochemical pathway is defined by Mavrovouniotis M.L. (1995) as an abstraction of a subset of intricate networks in the soup of interacting biomolecules. A prevailing definition is that a metabolic pathway is a special case of a metabolic network with distinct start and end points, initial and terminal vertices, respectively, and a unique path between them, i.e. a directed reaction graph with substrates as vertices and arcs denoting enzymatic reactions (Forst, Schulten, 1999). Typical metabolic pathways are given by the wall chart of Boehringer Mannheim [http://www.expasy.ch/tools/pathways/] and KEGG [http://www.genome.ad.jp/kegg/metabolism.html], which have been verified with a number of printed and on-line sources. Databases such as KEGG, WIT [http://

wit.mcs.anl.gov/WIT2/] represent metabolic pathway graphs with labeled arcs indicating the involved enzymes. Traditionally metabolic pathways have been defined in the context of their historical discovery, often named after key molecules (e.g. “glycolysis”, “urea cycle”, “pentose phosphate pathway” and “citric acid cycle” and so on). Schuster *et al.* (2000) provided a general definition of metabolic pathways based on the concept of elementary flux modes. The basic strategy to represent and compute pathways is the reactant-product binary relation. Properties of the pathway that rely upon the integration of two or more input molecules and unrelated output molecules and feedback effects are ignored.

Obviously, a metabolic pathway is a special part of complex network of reactants, products and enzymes with multiple interconnections representing reactions and regulation. One is called a pathway only if they are linear and non-branched. A pathway’s substrates are usually the products of another pathway, and there are junctions where pathways meet or cross. We consider that a metabolic pathway is a subset of reactions that describe the biochemical conversion of a given reactant to its desired end product.

Let $M = \{m_1,...,m_n\}$ be a set of metabolites in cells. Let $e_i : M \to M$ be a function for enzymatic reactions taking place in the cells.

The fact that e_i is a function from a set of substrates S ($S \subseteq M$) into a set of products P ($P \subseteq M$). It can be written as follows:

$$e_i : S \to P$$

for all m_1, m_2, $m_3 \in M$, the following property holds:

$$e_1(m_1) = m_2 \text{ and } e_2(m_2) = m_3 \Rightarrow e_2(e_1(m_1)) = m_3.$$

Let $e_1(m_1) = m_2$, $e_2(m_2) = m_3$, …., $e_k(m_k) = m_{k+1}$, we define $e_1e_2...e_k(m_1) = e_k(e_{k-1}...e_1(m_1)) = m_{k+1}$.

A new proposed definition of the metabolic pathway is presented and discussed in the following paragraphs.

Definition 1. *Given* $e_i : M \to M$, a *metabolic pathway is defined as a subset of successive enzymatic reaction events* $P = e_1e_2...e_k$.

Each enzymatic reaction e_i ($1 \le i \le k$) is catalyzed by a certain enzyme that is denoted as a unique EC number. The EC number is expressed with a 4-level hierarchical scheme that has been developing by the International Union of Biochemistry and Molecular Biology (IUBMB). The 4-digit EC number, $d_1.d_2.d_3.d_4$ represents a sub-sub-subclass indication of biochemical reaction. For instance, arginase is numbered by EC 3.5.3.1, which indicates that the enzyme is a hydrolase (EC 3.*.*.*), acts on the “carbon-nitrogen bonds, other than peptide bonds” (sub-class EC 3.5.*.*) in linear amidines (sub-sub-class EC 3.5.3.*). Thus we can adapt the EC number as a unique name for the responding enzyme catalyzed reaction.

Metabolic Pathway Alignment

Theory Basics

In order to score the similarity (percent identity) between two metabolic pathways, we define the similarity function.

Definition 2. *Let E be a finite set of e functions, an edit operation is an ordered pair*

$$(\alpha,\beta) \in (E \cup \{\varepsilon\}) \times (E \cup \{\varepsilon\}) \setminus \{(\varepsilon,\varepsilon)\} \cdot$$

α and β denote 4-digit EC strings of enzymatic reaction function, e.g. $\alpha = e_{1.1.1.1}$ $\beta = e_{2.3.4.5}$, ε denotes the empty string for null function. However, if $\alpha \neq \varepsilon$ and $\beta \neq \varepsilon$, then the edit operation (α,β) is identified with a pair of enzymatic reaction function.

An edit operation (α,β) is written as α→β (we can simply written α, β as EC numbers). There are three kinds of edit operations:

$\alpha \rightarrow \varepsilon$ denotes the deletion of the enzymatic reaction function α,

$\varepsilon \rightarrow \beta$ denotes the insertion of the enzymatic reaction function β,

$\alpha \rightarrow \beta$ denotes the replacement of the enzymatic reaction function α by the enzymatic reaction function β,

notice that $\varepsilon \rightarrow \varepsilon$ never happens.

Definition 3. *Let $E_1=e_1e_2...e_m$ and $E_2=e_1'e_2'...e_n'$ be two metabolic pathways, an alignment of E_1 and E_2 is a pair sequence*

$$(\alpha_1 \rightarrow \beta_1,\ldots,\alpha_h \rightarrow \beta_h)$$

of edit operations such that $E_1'=\alpha_1,...,\alpha_h$ and $E_2' = \beta_1,..., \beta_h$.

Example 1. *The alignment* A = (2.4.2.3→2.4.2.4, 3.5.4.5→ε, 3.1.3.5→3.1.3.5, ε→2.7.4.9) of the pathways $e_{2.4.2.3}e_{3.5.4.5}e_{3.1.3.5}$ and $e_{2.4.2.4}e_{3.1.3.5}e_{2.7.4.9}$ is written as follows, one over the other:

$$\begin{pmatrix} 2.4.2.3 & 3.5.4.5 & 3.1.3.5 & \varepsilon \\ 2.4.2.4 & \varepsilon & 3.1.3.5 & 2.7.4.9 \end{pmatrix}$$

Similarity Function

Definition 4. *A similarity function σ assigns to each edit operation* (α,β) *a nonnegative real number. The similarity* σ (α,ε) *and* σ (ε,β) *of the deletion operation* (α,ε) *and insertion operation* (ε,β) *is* 0. *For all replacement operations* (α,β) $\alpha\neq\varepsilon$, $\beta\neq\varepsilon$, *say, $\alpha = d_1.d_2.d_3.d_4$ and $\beta=d_1'.d_2'.d_3'.d_4'$, then the similarity function* $\sigma(\alpha,\beta)$ *is defined by*:

$$\sigma(\alpha,\beta) = \begin{cases} 0, \text{ if } (d_1 \neq d_1'); \\ 0.25, \text{ if } (d_1=d_1' \text{ and } d_2 \neq d_2'); \\ 0.5, \text{ if } (d_1=d_1' \text{ and } d_2=d_2' \text{ and } d_3 \neq d_3'); \\ 0.75, \text{ if } (d_1=d_1' \text{ and } d_2=d_2' \text{ and } d_3= d_3' \text{ and } d_4 \neq d_4'); \\ 1, \text{ if } (d_1=d_1' \text{ and } d_2=d_2' \text{ and } d_3= d_3' \text{ and } d_4 = d_4' \text{ i.e. } \alpha=\beta). \end{cases}$$

The definition does not exclude the possibility that d_4, $d_3.d_4$, and $d_2.d_3.d_4$ can be respectively expressed as wide card symbols *, *.* and *.*.* which means no clear classification of the enzyme.

However single pair of EC string comparison just means to measure how different EC strings are. Often it is additionally of interest to analyze the total difference between two strings into σ collection of individual elementary differences. The most important mode of such analyses is an alignment of the pathways. The function s can be extended to alignments in a straightforward way: the similarity σ (A) of an alignment $A=(\alpha_1 \rightarrow \beta_1,\ldots,\alpha_h \rightarrow \beta_h)$ is the sum of the similarities of the edit operations A consists of.

$$\sigma(A) = \sum_{i=1}^{h} \sigma(\alpha_i \rightarrow \beta_i).$$

A alignment scoring scheme, Score(E_1,E_2) of two metabolic pathways is the minimal mean similarity of their alignment

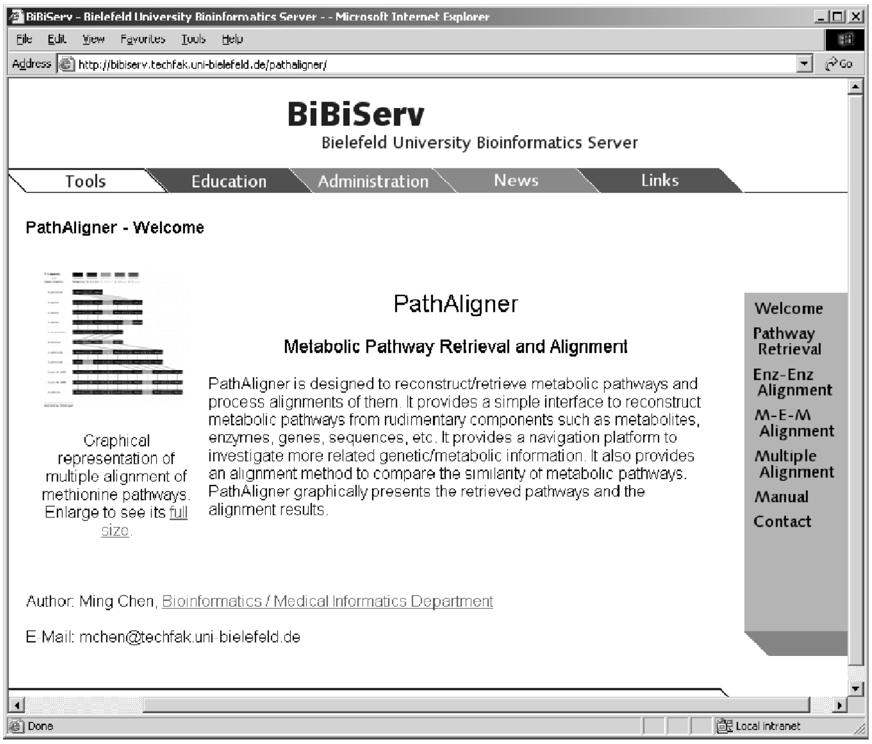
$$Score(E_1,E_2) = \frac{1}{max(m,n)}\sigma(A),$$

where m, n are the lengths of pathways.

Algorithms and Implementation

The pairwise alignment algorithm is as follows: 1. Initialize the set of unaligned EC number sequences, and the lengths; 2. Starting from both ends towards the middle, align one sequence to another and attempt to find all EC numbers with same 4-level hierarchical numbers. Score the similarities. Recall the alignment positions where EC number are identical and cut the sequences into more subsequences by removing the identical EC numbers; 3. Each pair of sub-sequences is initialized to begin a new round of 3-level hierarchical EC number matching till all pairs of sub-sequences are aligned. A similarity score is calculated afterwards; 4. Apply the same rule again, find the similarities of rest unaligned sub-sub-sequences based on 2-level hierarchical EC number matching and then sub-sub-sub-sequences on 1-level matching if any.

The algorithm has been implemented into the PathAligner system (http://bibiserv.techfak.uni-bielefeld.de/pathaligner) (Fig.).


BiBiServ - Bielefeld University Bioinformatics Server - - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://bibiserv.techfak.uni-bielefeld.de/pathaligner/
Go
BiBiServ
Bielefeld University Bioinformatics Server
Tools
Education
Administration
News
Links
Welcome
Pathway Retrieval
Enz-Enz Alignment
M-E-M Alignment
Multiple Alignment
Manual
Contact
Done
Local intranet

PathAligner - Welcome

PathAligner

Metabolic Pathway Retrieval and Alignment

Graphical representation of multiple alignment of methionine pathways. Enlarge to see its full size.

PathAligner is designed to reconstruct/retrieve metabolic pathways and process alignments of them. It provides a simple interface to reconstruct metabolic pathways from rudimentary components such as metabolites, enzymes, genes, sequences, etc. It provides a navigation platform to investigate more related genetic/metabolic information. It also provides an alignment method to compare the similarity of metabolic pathways. PathAligner graphically presents the retrieved pathways and the alignment results.

Author: Ming Chen, Bioinformatics / Medical Informatics Department

E-Mail: mchen@techfak.uni-bielefeld.de

Fig. A screenshot of PathAligner.

Three web-based alignment interfaces are implemented: "*E-E Alignment*", "*M-E-M Alignment*" and "*Multiple Alignment*". "*E-E Alignment*" uses the basic algorithm to align two linear metabolic pathways (represented as EC number sequences). User can also align any such a metabolic pathway against our pool database to find a list of hits. "*M-E-M Alignment*" considers the differences of metabolites in two pathways, which are presented as "Metabolite-EC number-Metabolite" patterns of sequence. It is possible to pick up two such pathways and align them to identify whether they are alternative pathways or partially are. "*Multiple Alignment*" allows the alignment of more than two metabolic pathways.

Conclusion

Identification and analysis of metabolic networks is a complex task due to the complexity of the metabolic system. Abstract pathway defined as a linear reaction sequence is practical for our alignment algorithm. We have presented an algorithm to study the problem of metabolic pathway alignment. Our algorithm calculates the hierarchical similarities of EC numbers mapping from both ends of the sequences. The algorithm has been successfully implemented into the PathAligner system.

References

Collado-Vides J., Hofestädt R. Gene regulation and Metabolism – Post genomic Computational Approaches. MIT Press, Cambridge, MA. 2002.

Dandekar T., Schuster S., Snel B. *et al*. Pathway alignment: application to the comparative analysis of glycolytic enzymes // Biochem J. 1999. V. 1. P. 115–24.

Forst C.V., Schulten K. Evolution of metabolisms: a new method for the comparison of metabolic pathways using genomics information // J. Comput. Biol. 1999. V. 6. P. 343–360.

Forst C.V., Schulten K. Phylogenetic analysis of metabolic pathways // J. Mol. Evol. 2001. V. 52. P. 471–489.

Mavrovouniotis M.L. Computational methods for complex metabolic systems: representation of multiple levels of detail // Bioinformatics & Genome Research / Eds. H.A. Lim , C.R. Cantor. World Scientific, 1995. P. 265–273.

Schuster S., Fell D., Dandekar T. A general definition of metabolic pathways uUseful for systematic organization and analysis of complex metabolic networks // Nature Biotechnol. 2000. V. 18. P. 326–332.

Tohsato Y., Matsuda H., Hashimot A. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy // Proc. 8th International Conference on Intelligent Systems for Molecular Biology (ISMB 2000), 2000. P. 376–383.

ABOUT NUMERICAL INVESTIGATION OF AUTO-OSCILLATIONS IN HYPOTHETICAL GENE NETWORKS

Kogai V.V.*[1], *Fadeev S.I.*[1], *Likhoshvai V.A.*[2,3]

[1] Sobolev Institute of Mathematics SB RAS, Novosibirsk, Russia; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [3] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia

* Corresponding author: e-mail: likho@bionet.nsc.ru

Keywords: *genetic systems, modeling, delay equation, differential autonomous systems*

Sammury

Motivation: The exposure of cycle operations in gene networks is an important problem of bioinformatics occupied with analyzing of structure and functioning regularities of gene networks and their regulator circuits.

Results: In the article we state the method of auto-oscillations numerical investigation described by differential autonomous systems of special form which model symmetrical canonical hypothetical gene networks of the first class (Likhoshvai *et al.*, 2003).

Introduction

Gene networks (GN) are structurally complicated spatial objects consist of dozens and hundreds elements of different nature and complexity: genes and regulatory sites; RNA and proteins coded by these genes; low-molecular compounds, different complexes between ferments and their targets etc. The core of GN is regulator circuits: genes and proteins whose expression is exposed to mutual regulation. Their presence provides GN for unique ability for adequate reaction on modification of environments. Thus, the exposure of possible functioning regimes of gene networks regulator circuits is an important problem of gene networks theory. There is a need to make systematical researches of functioning regularities of different constructions of gene networks regulator circuits to come to the solution of the problem. A design pitch in this direction is to pick up some finite set of standard elements from natural gene networks, to formalize rules of making up some theoretical objects from them (mathematical models) described gene networks regulator circuits, to carry out a systematical analysis of their properties for the purpose of the exposure of common biologically important regularities. Such works have been started by us in 2001 (Likhoshvai *et al.*, 2001). In this work we produce an effective seeking method of functioning cycle operations for symmetrical canonic hypothetical gene networks who are the simplest in design mathematical objects described in detail in article (Likhoshvai *et al.*, 2003). Biological meaning of this result consists in fact that this seeking method of functioning cycle operations is an essential stage of development of analysis methods for regulator circuits of arbitrarily structure.

Results

In this work we produce an effective numerical investigation method for auto-oscillations in autonomous differential equations sets described symmetrical canonic hypothetical gene network of the first class (Likhoshvai *et al.*, 2003):

$$dx_i / dt = \alpha/(1 + \beta \; z_i) - x_i, \; z_\mathrm{i} = \sum_{j=1}^{k-1} x_{\sigma(i-j)}^{\gamma}, \; \sigma(i-j) = \begin{matrix} i-j, \; if \quad j<i \\ n+i-j, \; if \quad j \geq i \end{matrix} \quad i = \overline{1,n}, \tag{1}$$

Where $\alpha > 0$, $\beta > 0$, $\gamma \geq 1$ are model parameters, $k \leq n$. According to (n,k) criterion formulated for symmetrical canonic HGN of the first class, there are $\bar{\alpha} > 0$ and $\bar{\gamma} > 1$ such that if $\alpha > \bar{\alpha},$ and $\gamma > \bar{\gamma}$ then autonomous system (1) has only k stable stationary solutions on conditions that greatest common divisor *d* of numbers *n* and *k* is equal to *k*. If $d \neq k$ then there are only *d* stable limit cycles and there is no stable stationary solution. Study of limit cycles parameters dependence is carried out on basis of the parameter continuation method with reference to boundary problem for the system (1) with boundary conditions represented periodic and transversal conditions (Kogai, Fadeev, 2001; Kogai, 2002):

$$0 \le t \le 1, \; dx_i/dt = T \; (\alpha/(1 + \beta \; z_i) - x_i), \; \mathrm{x_i}(0)=\mathrm{x_i}(1), \; \alpha \; (1 + \beta \; z_1) - x_1 = 0. \; i = \overline{1,n} \quad (2)$$

Where T is an oscillation period and is to be determined. It is important to note that numerical construction of the boundary problem (2) depending on parameter is not connected to stability of limit cycles and, thus, during the process of parameter continuation we have a possibility to find both stable and unstable limit cycles. If there was a need we determined the stability character using algorithm of maximum eigenvalue calculation of the monodromy matrix. We used a limit cycle of the system (1) found at some value of α as an initial solution of the system in the parameter continuation method.

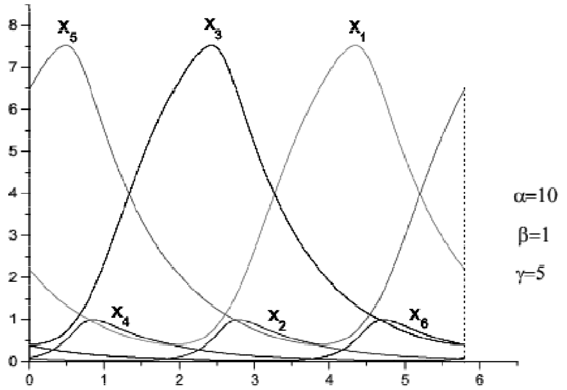


Fig. 1. Periodic solution of the $M_1(6,4)$ model with 2 groups of components.

Structure peculiarities of periodic solutions of $M_1(n,k)$ model allow to organize numerical investigations of the boundary problem (2) recruiting equations with lagging arguments. The method is based on numerical observation which shows in turn that limit cycles obtained by means of integration of the system (1) have partial symmetry property. In other words all components of limit cycles split up into *d* groups: $u_j=\{x_j+ld| \, l=0,\ldots,n/d-1\}$, j=1,…,*d*; and variables have the same amplitude and differ only in phase in each group (Fig. 1).

Thus, we come to the equivalent presentation of the boundary problem (2) in the form of boundary problem having *d* equations with lagging arguments:

$$0 \le t \le 1, \quad du_j/dt = T \;\; f_j(u(t), \; u(t-\tau_1),\ldots, \quad u(t-\tau_m)) \;,$$

$$\mathrm{u_j}(0)=u_j(1), \quad j=\overline{1,d} \;\; du_1(0)/dt = 0. \qquad (3)$$

Where u is a vector argument with components $u_1,\ldots,u_d$. Number and values of delays τ_i, connections with components of the boundary problem (2) are determined on base of the initial solution graphs and then by means of direct substitution in the equations set (1) we check the consistency of obtained presentation.

On Fig. 2 an example of numerical investigation is presented. In this example we consider auto-oscillations of the $M_1(6,4)$ model. At that the results for the boundary problem (2) are congruent the results getting by means of numerical investigation program of the boundary problem (3). For the $M_1(6,4)$ model the boundary problem (3) is formulated in the following way:

$$0 \le t \le 1, \quad du_1(t)/dt = T[\alpha/(1+\beta(u_2^\gamma(t-1/3)+u_1^\gamma(t-1/3)+u_2^\gamma(t-2/3)) - u_1(t)]$$

$$du_2(t)/dt = T[\alpha/(1+\beta(u_1^\gamma(t)+u_2^\gamma(t-1/3)+u_1^\gamma(t-1/3)) - u_2(t)] \qquad (4)$$

$$u_1(0)=u_1(1), \quad u_2(0)=u_2(1), \quad du_1(0)/dt = 0.$$

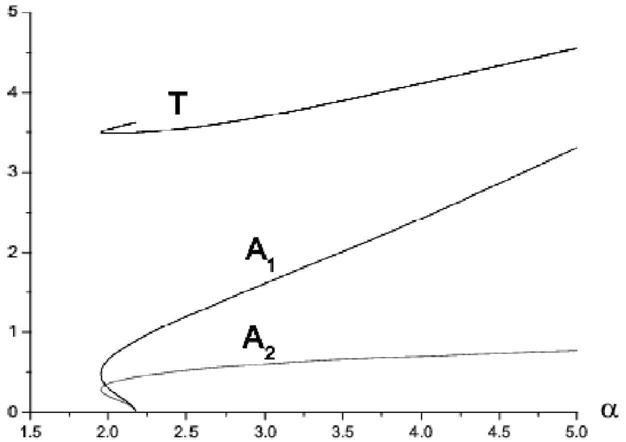


Fig. 2. Parameter α dependence of the period T and amplitude of oscillation A_1 and A_2 in the part symmetric periodic solution of the $M_1(6,4)$ model when β = 1, γ = 5.

The boundary problem (4) allow to find 2 cycles in the $M_1(6,4)$ model at once. We get the first cycle when the components $x_1(t)$, $x_3(t)$ and $x_5(t)$ belong to the 1 group and the components $x_2(t)$, $x_4(t)$ and $x_6(t)$ belong to the 2 group. And we get the second cycle when the components $x_2(t)$, $x_4(t)$ and $x_5(t)$ belong to the 1 group and the components $x_1(t)$, $x_3(t)$ and $x_5(t)$ belong to the 2 group.

Next generalization of described procedure allow to find all partly symmetrical cycles of model (1). In this case, the generalization lies in assumption that upon partitioning n variables on d groups, d is the denominator of *n* and *k*. From the same considerations it follows that there is natural number *s* such that for each variable *j+isd* belonged to the *j*-group we have the following equality:

$x_j(t) = x_{\mathrm{mod}_n(j+isd)}(t-i\tau)$, $i = \overline{0, n/d-2}$, $n\tau/d = T$, where T is a period. In order to find all the cycles of the system (1) it is necessary to establish by direct checking all consistent *d* and *s* and then solve for them proper problems (2). If there is a cycle in (2), then there is a cycle in (1) as well.

So, in the considered model there is a possibility to have “symmetrical” cycle for which d=1, s=5, τ=T/6 (Fig. 3). To seeking this cycle the boundary problem for the equation with lagging argument has the next form:

$$0 \le t \le 1, \quad du_1(t)/dt = T[\alpha/1+\beta(u_1^\gamma(t-1/6)+u_1^\gamma(t-1/3)+u_1^\gamma(t-1/2)) - u_1(t)]$$

$$u_1(0)=u_1(1), \quad du_1(0)/dt = 0. \tag{5}$$

Except for stability problem, an effectiveness of offered method is obvious since the boundary problem (2), having *n* equations, is formulated now in form having only *d* equations with lagging arguments.

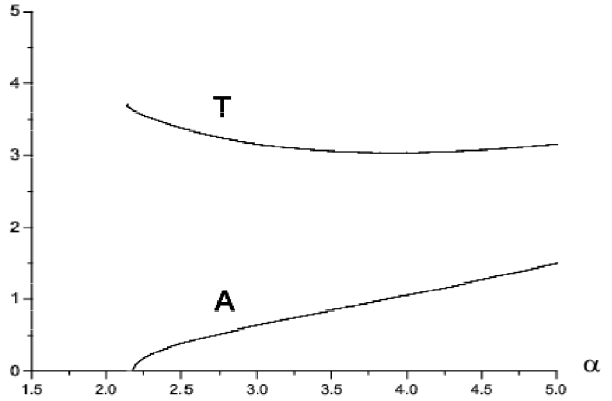


Fig. 3. Parameter α dependence of the period T and amplitude of oscillation A of the symmetric periodic solution of the $M_1(6,4)$ model when β = 1, γ = 5.

Conclusion

One of the dominant problems of gene networks functioning dynamic investigation is the problem of search of oscillation regimes. Problem not of the less importance is the problem of exposure of gene networks structure regularities which are in charge of formation of cycle operation. This article has a methodical character and was made within the bounds of general line which goes in for researches of gene networks functioning regularities. In this article we offer a new effective method of search of symmetrical canonic hypothetical gene networks cycle operations. In the future we plan to use obtained result for the development of more general method of exposure of oscillation functioning regimes of arbitrary structure gene networks.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (No. 02-04-48802, 02-07-90359, 03-01-00328, 03-04-48506, 03-04-48829, 03-07-96833, 04-01-00458), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4) and "Origin and evolution of biosphere" (No. 10002-251/П -25/ 155-270/200404-082); SB RAS (Integration Project No. 119).

References

Kogai V.V., Fadeev S.I. Application of parameter continuation on base of multiple shooting method for numerical investigation of nonlinear boundary problems // Sib. J. of Industrial Mathematics. 2001. V. 4. P. 83–101.

Kogai V.V. Application of parameter continuation for numerical investigation of periodic solutions of ordinary differential autonomous systems // NSU Messenger. 2002. V. 2. P. 40–48.

Likhosvai V.A., Fadeev S.I. About hypothetical gene networks // Sib. J. of Industrial Mathematics. 2003. V. (15). P. 134–153.

Likhosvai V.A., Matushkin Yu.G., Fadeev S.I. Problems of functioning theory of gene networks // Sib. J. of Industrial Mathematics. 2003. V. 4. P. 64–80.

BIOUML – OPEN SOURCE EXTENSIBLE WORKBENCH FOR SYSTEMS BIOLOGY

Kolpakov F.A.

Biosoft.Ru/DevelopmentOnTheEdge.com, Novosibirsk, Russia;
Design Technological Institute of Digital Techniques SB RAS, Novosibirsk, Russia,
e-mail: fedor@biosoft.ru

Keywords: *systems biology, database, visual modeling, plug-in based architecture, Java, Eclipse*

Summary

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. This step requires integrated software environment that spans the comprehensive range of capabilities.

Results: BioUML – Biological Universal Modeling Language – is open source extensible Java workbench for systems biology. BioUML's core is a meta model that provides an abstract layer for comprehensive formal description of wide range of biological and other complex systems. Content of databases on biological pathways (TRANSPATH, KEGG/pathways, GeneNet, GeneOntology) as well as SBML and CellML models can be expressed in terms of the meta model and used by BioUML workbench. Plug-in based architecture provides the workbench extensibility and possibility of seamless integration with other tools for systems biology. The workbench consists from Eclipse platform runtime that supports plug-ins registry and a set of plug-ins for database access, diagram editing, biological systems simulation and for integration with MATLAB and SBW/SBML.

Availability: http://www.biouml.org

Introduction

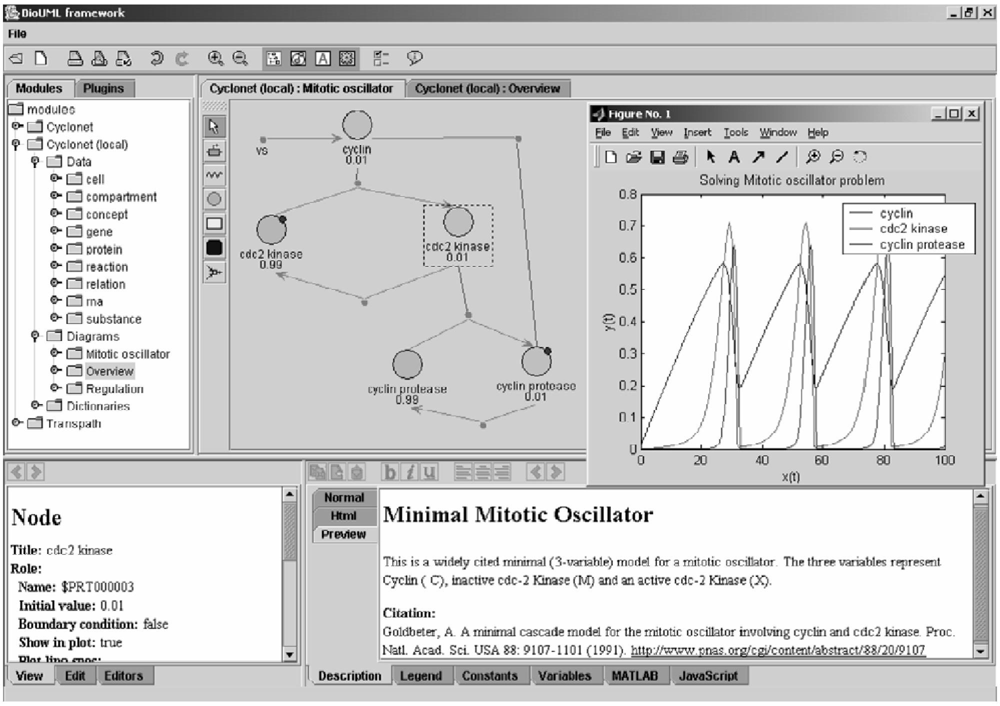
Sydney Brenner, 2002 Nobel Prize winner said (Bradford, 2003):

"*We now have unprecedented ability to collect data about nature but there is now a crisis developing in biology, in that completely unstructured information does not enhance understanding. We need a framework to put all of this knowledge and data into — that is going to be the problem in biology. We've reached the stage where we can't talk to each other — we've all become highly specialized. We need a framework, a framework where people can come back to us and say, 'Yes, I understand.' Driving toward that frame-work is really the big challenge.*"

We believe that BioUML – Biological Universal Modeling Language – is a step in this direction. It is imagined as a language to write a "book of life". From the user's perspective BioUML workbench (Fig. 1) is integrated 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.

Architecture overview

BioUML workbench is a plugin-based application framework (Fig. 2) based on Eclipse platform runtime (IBM, 2003). It consists from a core runtime that supports 'plug-ins' and a set of plug-ins that support database access, diagram editing, simulation, integration with MATLAB and SBW/ SBML (Hucka *et al.*, 2002, 2003), etc.



Node

Title: cdc2 kinase
Role:
Name: $PRT000003
Initial value: 0.01
Boundary condition: false
Show in plot: true
Plot line spec:

Minimal Mitotic Oscillator

This is a widely cited minimal (3-variable) model for a mitotic oscillator. The three variables represent Cyclin (C), inactive cdc-2 Kinase (M) and an active cdc-2 Kinase (X).

Citation:
Goldbeter, A. A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. Proc. Natl. Acad. Sci. USA 88: 9107-1101 (1991). http://www.pnas.org/cgi/content/abstract/88/20/9107

Fig. 1. BioUML workbench screenshot. Top left pane – repository pane that shows database modules, here Cyclonet database module; top right pane – diagram editor; bottom left pane – property editor for the selected object; bottom right pane – diagram editor parts, here the diagram description editor; middle right pane – results of the simulation.

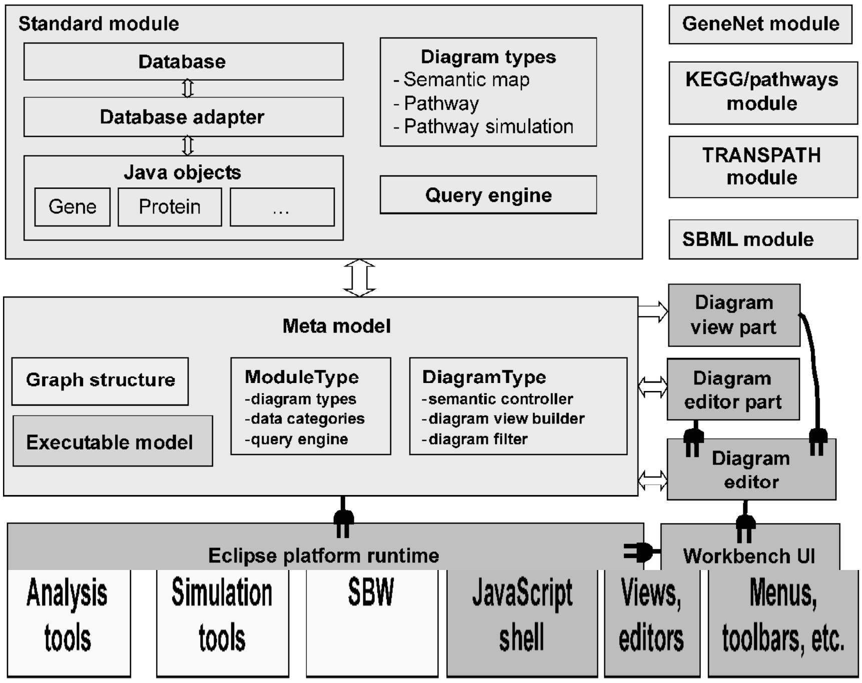


Fig. 2. BioUML workbench architecture overview.

A plug-in is the smallest unit of BioUML workbench function that can be developed and delivered separately into BioUML workbench. Extension points are well-defined function points in the system where other plug-ins can contribute functionality. An extension is a specific contribution to an extension point. Plug-ins can define their own extension points, so that other plug-ins can integrate tightly with them.

Plug-ins are coded in Java. A typical plug-in consists of Java code in a JAR library, some read-only files, and other resources such as images, native code libraries, etc. BioUML workbench installation includes a plug-ins folder where individual plug-ins are deployed. Each plug-in is installed in its own folder. A plug-in is described in an XML manifest file, called plugin.xml, residing in the plug-in's folder. The parsed contents of plug-in manifest files are made available programmatically through a plug-in registry API provided by Eclipse runtime.

Meta model

The core of BioUML workbench is meta model that provides an abstract layer for comprehensive formal description of wide range of biological and other complex systems. Content of databases on biological pathways as well as CellML and SBML models can be expressed in terms of meta model and then can be visualized and edited as diagram by BioUML diagram editor, simulated using MATLAB or BioUML simulation engine, etc.

Meta model is problem domain neutral and splits the system description into three interconnected levels:

1. graph structure – the system structure is described as compartmentalized graph;

2. database level – each graph element can contain reference to some database object;

3. mathematical model – any graph element can be element of executable (mathematical) model.

Figure 3 demonstrates how this approach is applied for modeling system consisting from two consecutive chemical reactions. Here graph nodes representing chemical substances are considered as variables and corresponding graph edges contain right parts of corresponding differential equations. Using this information BioUML workbench can generate MATLAB or Java code for model simulation.

Discussion

Formal description and modeling of biological systems require coordinated efforts of different group of researchers:

1) programmers – they should provide computer tools for this task;

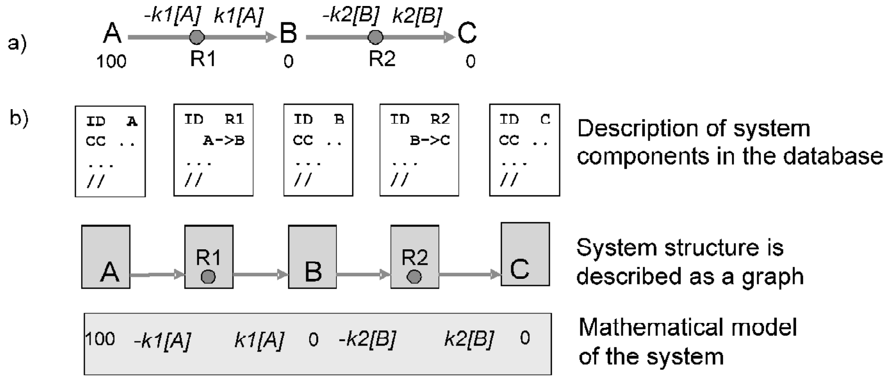
2) problem domain experts – they should specify what and how should be described;

3) experimenters and annotators – they should describe corresponding data following to these rules;

4) mathematicians – they should provide methods for models analysis and simulations.

I believe that one of the methodological achievements of BioUML workbench is that it separates these tasks so they can be effectively solved by corresponding group of researchers and provides simple contract how these groups (and corresponding software parts) should communicate one with other. Meta-model is the base of this contract for all parties.

Plug-in based architecture provides BioUML workbench extensibility and possibility of seamless integration of other tools for systems biology. Freely available BioUML workbench source code allows customers to develop their own plug-ins and database modules to extend BioUML workbench for their needs. For this purpose there is special BioUML development kit that includes all source code and all needed third party libraries (http://www.biouml.org).


a)
-k1[A] k1[A]
-k2[B] k2[B]
A
B
C
100
R1
0
R2
0
b)
ID A
CC ..
...
//
ID R1
A->B
...
//
ID B
CC ..
...
//
ID R2
B->C
...
//
ID C
CC ..
...
//
Description of system components in the database
A
R1
B
R2
C
System structure is described as a graph
100 -k1[A] k1[A] 0 -k2[B] k2[B] 0
Mathematical model of the system

c)

$$\frac{d[A]}{dt} = -k_1[A]$$

$$\frac{d[B]}{dt} = k_1[A] - k_2[B]$$

$$\frac{d[C]}{dt} = k_2[B]$$

$[A](t_0) = 100$ mM/l
$[B](t_0) = 0$ mM/l
$[C](t_0) = 0$ mM/l

Fig. 3. System from two consecutive chemical reactions (a), its formal description using three meta model levels (b), and corresponding mathematical model (c), that can be generated automatically for system simulations.

Acknowledgements

The work was supported by the grants INTAS Nr. 03-51-5218 and RFBR Nr. 04-04-49826-a. The author is grateful to Biosoft.Ru/DevelopmentOnTheEdge.com stuff for technical support.

References

Bradford R. A man, a worm, and A nobel // Salk Signals. 2003. V. 5(2). P. 12–17.

Hucka M., Finney A., Sauro H.M., Bolouri H. *et al.* The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models // Bioinformatics. 2003. V. 19(4). P. 524–531.

Hucka M., Finney A., Sauro H.M., Bolouri H., Doyle J., Kitano H. The ERATO systems biology workbench: enabling interaction and exchange between software tools for computational biology // Proc. of the Pacific Symposium on Biocomputing 2002.

IBM. Eclipse platform. 2003. http://www.eclipse.org

EVOLUTIONARY ALGORITHMS FOR MATHEMATICAL MODELS OF GENE NETWORKS IDENTIFICATION

Lashin S.A. [1]*, *Likhoshvai V.A.*[1,2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Corresponding author: e-mail: lashin@bionet.nsc.ru

Keywords: *genetic algorithm, evolutionary algorithm, parameter identification, inverse problem, gene network*

Summary

Motivation: Mathematical models of biological processes possessing giant dimension (thousands and more variables) is the prime interest of mathematical biology. It refers, in particular, to the models of gene networks. A key stage of adequate mathematical models of gene networks construction is the stage of identification of their parameters (so called inverse problem). For the successful solution of the given problem availability of a significant array of the quantitative dynamic data on functioning modeled genetic systems is necessary. Nowadays, due to new microchip biotechnologies development, there is a prompt accumulation of simultaneous *in situ* data under quantitative characteristics of biological systems functioning dynamics. It makes a considered problem be the one of the most actual problems of mathematical biology.

Results: We have developed the set of methods for identification of the giant dimension models (thousand and tens thousand variables) – i.e. finding of model parameter values at which the level of plausible conformity of model calculations results with experimental data is reached. The methods are based on ideas of evolutionary and genetic algorithms of optimization. Also program realizations of methods, including their parallel versions in MPI standard are developed.

Introduction

The problem of mathematical model parameter finding by its behavior is called as identification of mathematical model parameters (inverse problem). All methods of parameter identification can be divided in two classes. The first one uses the model calculation during iterative process (direct problem), the other dispenses this process.

The essential part of the first type methods is so-called evolutionary and genetic methods known for the application in problems of optimization (Batischev, 1995). For using this methods in a problem of identification, it is represented as optimization one where the "distance" between the model dynamics calculated with the given parameters and experimental dynamics serves as criterion function. This target setting of inverse problem directly leads it to be solved with the use of evolutionary and genetic methods.

Algorithms

The idea of algorithm consists in consecutive evolution of so-called "individuals" – models of the gene networks possessing "genotype" – a concrete set of model constants. The genotype unambiguously determines "phenotype" – dynamics of a gene network. Individuals form "population" which is a material for evolutionary selection of the most adapted individuals. Evolution is a consecutive change of structure of a population which is carried out for one evolutionary step. The evolutionary step is an elementary unit of evolutionary process and, in turn, is composed from the following substeps.

The first substep consists in selection of a subpopulation which members generate descendants. Selection of a subpopulation is defined by a separate model which can be modified as the separate module. At present time, a number of models of a subpopulation selection are developed. The first model suggests the fixed quantity of the individuals having the best characteristics of individual fitness to be selected for the subsequent reproduction. In the other model probabilistic selection where the probability to take part during reproduction depends definitely on fitness of an individual (more adapted individuals have more chances to continue the family, however the weakest individuals also are not deprived such opportunity, as against the model considered above) is realized. In this model the number of individuals in a subpopulation can change on different steps of algorithm. All other individuals which have not got in a subpopulation are considered dyed out and further are not reviewed.

The second substep consists in reproduction of individuals from the selected subpopulation. This substep also is carried out on the basis of separate model. At present time there are some models of reproduction. This process can be carried out by different ways – each survived individual can give strictly fixed or any (limited only by number of descendants) quantity of descendants. After the process completion there will be a population containing the parent survived individuals, and also their descendants. Reproduction is implemented under the control of the separate module which can contain models of mutational process. Several models of "mutations" generation are developed. For example, one of models generates affiliated individuals which differ from parent no more than in one parameter (model of a dot mutation). In the other model the maximal allowable number of mutations on one individual is fixed, and affiliated individuals can have some differences in constants, but not more than the fixed number (this model realizes Haldane's dilemma – evolutionary process cannot go on the big number of parameters simultaneously). Except for models of mutations generation the models realizing sexual reproduction (crossingover models) were considered. In these models the affiliated individual receives a part of parameters from one parent individual, the rest of parameters – from the second parent individual. Models can differ in crossingover type (simple or plural) and quantity of parents of an affiliated individual (from two to all individuals in a subpopulation).

When the population generation process is completed, for each individual the phenotype (dynamics of an individual functioning) is calculated. Phenotype found compares with a sample phenotype - experimental data about this dynamics after that fitness function can be calculated. Thus the third substep consists in fitness function calculation. Function of fitness also is model and can be represented as mean square deviation, weighed mean square deviation and other distances in the space of (discrete) functions.

The received results move by the beginning of the following evolutionary step, process of selection and reproduction repeats.

As a result of such selection individuals in a population become gradually closer and closer on a phenotype to the sample. Process of evolution proceeds until one or several individuals will not have a phenotype that is close enough to the sample. After that genotypes of these best individuals can be considered as solutions of inverse problem. Now the given method is successfully applied to search of optimum sets of parameters of mathematical models of the gene networks, which are listed above.

At present the number of individuals in a population is constant and equal to numbers of nodes (processes) of the parallel program.

Realization and Results

The algorithms described above are realized as a package in C ++ language, with use of library of parallel calculations MPI (Snir *et al.*, 1996; Korneev, 2003). Classes of an individual and a population, and also models of genetic algorithms are realized. Due to use of the object-oriented approach by development, the package easily extends and changes.

The package is applied to solve inverse problem of dynamic models of gene networks functioning. Models are described in terms of elementary processes - biochemical reactions. For calculation of dynamics of models (direct problem solution) the software package based on GCKM (the generalized chemical-kinetic method) (Likhoshvai *et al.*, 2001) is used.

Problems are calculated on high-efficiency computers, such as MVS-1000, SunFire 15K and others.

Problem solution of the test dynamic model identification is suggested to demonstrate a parallel approach advantage. The offered model is the simple biological oscillator described by the following ordinary differential equations

$$\begin{cases} \dfrac{dX_1}{dt} = \dfrac{\alpha_1}{1 + X_3^{\gamma 1}} - k_{d1} X_1 \\ \dfrac{dX_2}{dt} = \dfrac{\alpha_2}{1 + X_1^{\gamma 2}} - k_{d2} X_2 \\ \dfrac{dX_3}{dt} = \dfrac{\alpha_3}{1 + X_2^{\gamma 3}} - k_{d3} X_3 \end{cases}$$

Model was calculated on time interval [0,200]. The sample dynamics was obtained by calculation of the model with the constants (α_1=1, α_2=2, α_3=3, γ_1=2, γ_2=4, γ_3=2, k_{d1}=0.004, k_{d2}=0.002, k_{d3}=0.001) (Fig. a, b, c, curves with number 1). The evolutionary process had started from the following set of constants (α_1=1.5, α_2=2.5, α_3=3, γ_1=2.5, γ_2=4, γ_3=1.5, k_{d1}=0.008, k_{d2}=0.002, k_{d3}=0.003). That dynamics is shown in Fig. a, b, c (curves with number 3). The dynamics of the model specified by inverse problem solution also is given in Fig. a, b, c (curves with number 2).

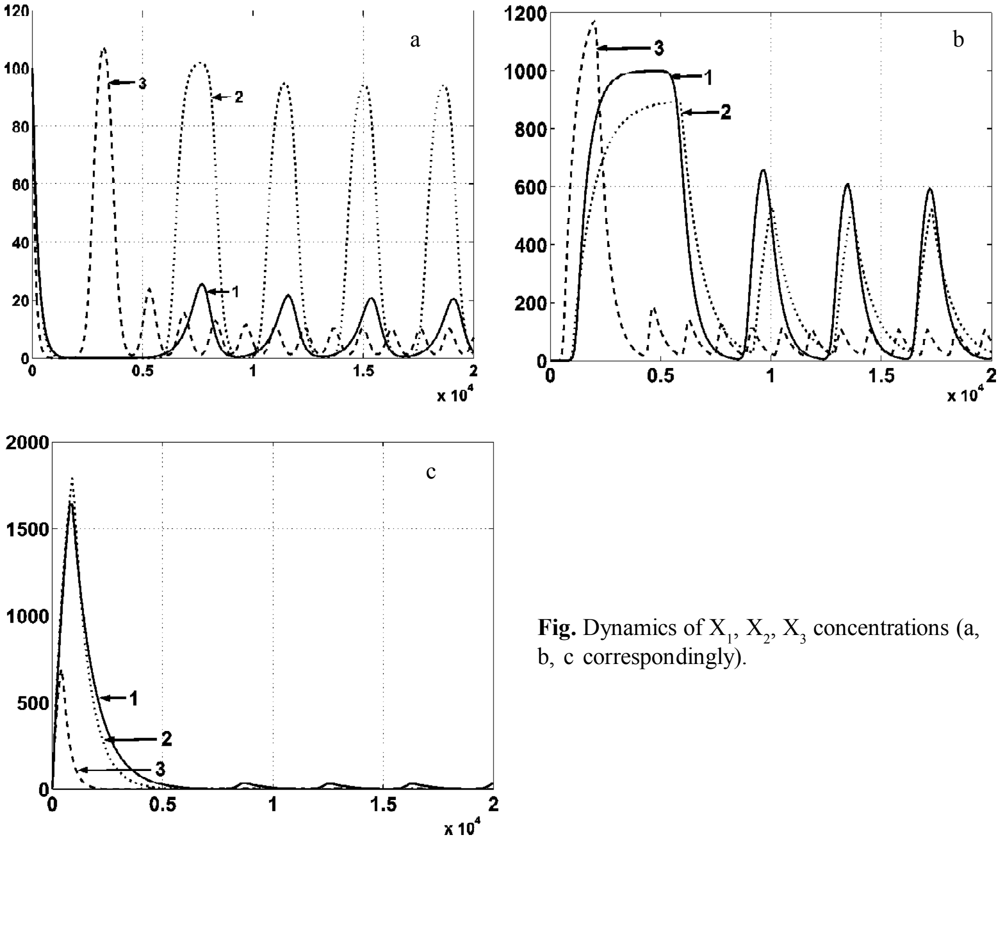


Fig. Dynamics of X_1, X_2, X_3 concentrations (a, b, c correspondingly).

The problem was solved with the use of different package options. According to maximal number of individuals in the population the number of iterations varied (Table). Also the times expended to get solutions (with parallelization switched off/on) obviously show the advantage of the parallel version of the method.

Table. Time and number of iterations needed to inverse problem solution

Number of individuals in population	Parallelization	
	switched off	switched on
10	iterations 62, time 2480 sec.	iterations 62, time 280 sec.
20	iterations 58, time 4640 sec.	iterations 58, time 258 sec.
40	iterations 49, time 7840 sec.	iterations 62, time 220 sec.
80	iterations 40, time 12800 sec.	iterations 40, time 190 sec.

Two stages of evolutionary process were found. On the first one the individuals of population improve their fitness rapidly. On the other stage average population fitness starts to oscillate and the further adaptation is too slow. It should be noted that even first stage of the evolution process has biological meaning (because of biological data is often imperfect).

The programs realized allow applying the methods obtained to the mathematical models of the giant dimension identification. Due to directivity on parallel calculations, the significant gain in speed is reached at the solution of this problem.

Acknowledgements

The work was supported by the leading scientific schools grant No. 311.2003.1 of President of Russia, by the grants No. 04-01-00458, 03-07-96833, 03-04-48506, 03-01-00328, 02-07-90359, 02-04-48802 of the RFBR, by the interdisciplinary grant No 119, Project No. 10.4 of the RAS Presidium Program "Molecular and Cellular Biology"; and Program for Basic Research of the Presidium of the Russian Academy of Sciences (contract No. 10002-251/П-25/155-270/200404-082), Integration project of the RAS (No. 148).

References

1. Batischev D.I. Genetic algorithms for solving extreme problems / Eds. Y.E. Lvovich. Voronezh, 1995. (In Russ.).
2. Korneev V.D. Parallel programming in MPI. / Eds. V.E. Malyshkin, O.L. Bandman. Institute of computer research, Moscow- Izhevsk, 2003. (In Russ.)
3. Likhoshvai V.A., Matushkin Iu.G., Ratushnyi A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. A generalized chemical-kinetic method for modeling gene networks // Mol. Biol. (Mosk.). 2001. V. 35(6). P. 1072–9. (In Russ.).
4. Snir M., Otto S., Huss-Lederman S., Walker D., Dongarra J. MPI: The Complete Reference. MIT Press, Boston, 1996.

EXPLICIT INTEGRAL METHOD FOR NONLINEAR DYNAMIC MATHEMATICAL MODELS IDENTIFICATION

Lashin S.A.[1]*, *Likhoshvai V.A.*[1,2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Corresponding author: e-mail: lashin@bionet.nsc.ru

Keywords: *explicit method, parameter identification, inverse problem*

Summary

Motivation: One of the basic stages the mathematical models construction of biological processes is the identification stage of their parameters. Most of the methods focused on this problem uses process of calculation of model (numerical integration of system), that strongly increases computing expenses at the solving. The explicit method of parameter identification, dispensing integration of system and, thereof, demanding less computing expenses is offered. It does a considered problem be an actual problem of the numerical analysis and the mathematical biology.

Results: The explicit method for high dimension dynamic models identification (finding of model parameter values at which the plausible conformity of model calculation results level with experimental data is reached) is developed. The method is based on integral transformation of initial system of the differential equations, approximation of discrete values of the model functions with the experimental data help, and the following solution of linear systems of the equations. Program realization of the method, including the parallel version in MPI standard is developed.

Introduction

Consider a class of the nonlinear dynamic models represented by a system of the differential equations in explicit Cauchy's form:

$$\frac{dX_i(t)}{dt} = f_i(C, X), \qquad (1)$$

where $X_i(t)$ – is the set of dynamic variables of studying phenomena ($i=1,...,N$), C – vector of model parameters.

The problem of parameters definition for mathematical model in its behavior is called as identification of mathematical model parameters (inverse problem). All methods of parameters identification can be separated on two classes. The first one uses the model calculation during iterative (direct problem), the other dispenses this process. Complexity of the first type methods, it is especial with increase in dimension of system, grows exponentially, since the complexity of the direct problem solution required on each iteration (integration of system) is growing,

The second type algorithms are based on the certain calculations, allowing avoiding the system numerical integration mentioned above. By a principle of realization of this idea the methods can be divided into two principal groups: differential and integral (Ermakova, 1989).

Differential methods approximate the values of derivatives dX_i/dt and right part $f_i(C,X)$ using the experimental data. After that the identification problem is reduced to a set of systems of the algebraic equations (Karnaukhov, Karnaukhova, 2003).

Integral methods are alternative to differential methods. They are based on integral transformations of initial system of the differential equations (1), with the following approximate calculation of integrals. In contrast to the differential methods, lacking that the approximate derivatives calculation is ill-conditioned problem, integral methods work with much smaller mistakes, since the approximate calculation of integrals is well-conditioned problem.

The method developed is integral and being designed for the identification problem solution of the certain kind systems (systems with a quadratically-linear right part), reduces an initial problem to a problem of the solution of a set of systems of the linear algebraic equations.

Algorithm

Consider a class of nonlinear dynamic models more particular than (1), namely, class of the systems of the differential equations with a quadratically-linear right part:

$$\frac{dX_i(t)}{dt} = \sum_{k \le l}^{N} c_{kl}^{i} X_k(t) X_l(t), \tag{2}$$

where $X_i(t)$ – is the set of dynamic variables of studying phenomena ($i=1,...,N$), c^i_{kl} – parameters (constants) of model ($k \le l=1,...,N$). If we integrate system (2) in the range of t_0 to t_m $(m=1,...,M)$, we will get:

$$X_i(t_m) - X_i(t_0) = \int_{t_0}^{t_m} \left(\sum_{k \le l}^{N} c_{kl}^{i} X_k(t) X_l(t) \right) dt = \sum_{k \le l}^{N} c_{kl}^{i} \int_{t_0}^{t_m} X_k(t) X_l(t) dt\,. \tag{3}$$

Thus, having enough of experimental data, we can calculate appropriate values

$$d_m^i = X_i(t_m) - X_i(t_0), \tag{4}$$

and

$$A_{klm}^i = \int_{t_0}^{t_m} X_k(t) X_l(t) dt\,. \tag{5}$$

Using notation (4), (5), expression (3) can be transcribed as

$$\sum_{k \le l}^{N} c_{kl}^{i} A_{klm}^{i} = d_m^i, \quad m=1,...,M \tag{6}$$

expression (6) can reduced to the standard form of the linear equations systems notation. We have to renumber pair indexes kl into single index j: $k \otimes l \leftrightarrow j$ $(k \le 1=1,...,N \leftrightarrow j=1,...,N(N+1)/2)$. After renumbering we can get linear system of algebraic equations in the standard notation form

$$\sum_{j=1}^{N(N+1)/2} c_j^i A_{jm}^i = d_m^i, \qquad m=1,...,M, \tag{7}$$

where parameters c^i_j are unknown quantities. On the assumption of $M=N(N+1)/2$ matrix A is square and the linear system obtained (7) can be solved using different numeric methods. Having solved the system, we shall find the parameters included in the i-th equation of system (2). Thus, having solved N systems, we shall find all parameters of system (2) and the problem of identification will be solved.

Realization and Results

The algorithm described above is realized using C ++ using the library of parallel calculations MPI (Snir *et al*., 2000; Korneev, 2003) and the libraries of linear algebra subroutines LAPACK (Blackford, 2000) and BLACS (Dongarra *et al.*, 1997; Whaley, 1997).

The program realized allows to apply the received methods to identification of mathematical models with more complex right part. Because of orienting on parallel calculations, the significant acceleration is reached at the salvation of the problem.

In our opinion the main advantage of the suggested algorithm that it is not excessing and comes nearer to an analytical way of the solution of identification problem. At enough of the data the method can be applied to identification of the structurally functional organization of gene networks.

Acknowledgements

The work was supported by the leading scientific schools grant No. 311.2003.1 of President of Russia, by the grant Nos. 04-01-00458, 03-07-96833, 03-04-48506, 03-01-00328, 02-07-90359, 02-04-48802 of the RFBR, by the interdisciplinary grant No 119, Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology” (Program for Basic Research of the Presidium of the RAS, contract No. 10002-251/ П -25/155-270/200404-082). The authors are indebted to A.V. Ratushny for technical assistance.

References

Blackford L.S., Choi J., Cleary A., D’Azevedo E., Demmel J., Dhillon I., Dongarra J., Hammarling S., Henry G., Petitet A., Stanley K., Walker D., Whaley R.C. ScaLAPACK Users’ Guide. 1997.

Dongarra J., Whaley R.C. A User’s Guide to the BLACS v1.1. 1997.

Ermakova A. New complex of numerical methods for identification and analysis of kinetic models problem // Mathematical modeling of catalytic reactors. Novosibirsk: Science. Sib. Branch, 1989. P. 120–150.

Karnaukhov A.V., Karnaukhova E.V. Use of new identification method for nonlinear dynamic systems for biochemistry problems // Biochemistry. 2003. V. 68(3). P. 309–317. (Russ.).

Korneev V.D. Parallel programming in MPI. / Eds. V.E. Malyshkin, O.L. Bandman. Institute of computer research, Moscow-Izhevsk, 2003. (Russ.).

Snir M., Otto S., Huss-Lederman S., Walker D., Dongarra J. MPI: The Complete Reference. MIT Press, Boston, 1996.

Whaley R.C. Outstanding Issues in the MPIBLACS. 1997.

EFFICIENT ALGORITHM FOR GENE SELECTION USING PLS-RLSC

Li Shen*, Eng Chong Tan

School of Computer Engineering, Nanyang Technological University,
Nanyang Avenue, Singapore 639798, Singapore
* Corresponding author: e-mail: shenli@pmail.ntu.edu.sg

Keywords: *cancer classification, gene selection, DNA microarray, support vector machines, partial least squares, regularized least squares classification, recursive feature elimination*

Summary

Motivation: Accurate cancer diagnosis is very important for treatment of cancer patients. Gene selection is crucial to classifier design for cancer classification using microarray data. Efficient and effective algorithms for cancer classification and gene selection are needed in this area.

Results: A new method called PLS-RLSC for cancer classification and gene selection is proposed. It is based on the partial least squares (PLS) as dimension reduction followed by regularized least-squares classification (RLSC). The new method performed empirically better than support vector machine (SVM) on the publicly available colon cancer dataset and required much less time. It is also combined with the recursive feature elimination (RFE) algorithm to select a six-gene subset to achieve the minimum testing errors. The testing accuracy is as high as 98 %.

Availability: The MATLAB source codes are available on request.

Introduction

The objective of cancer classification is to design a classifier to categorize the tissue samples into pre-defined classes (e.g. tumor and normal) using the gene expression levels produced by microarray techniques. Since the data dimension is very large, SVMs have been found to be very useful for this classification problem [2]. Apart from the classification task, it is also important to eliminate the irrelevant genes from the dataset and select a small subset of marker genes, which discriminate between the different types of tissue samples. Some techniques like RFE was proposed by other researchers to accomplish this task [5]. The RLSC is shown to be as good as SVM on several benchmark datasets [6]. We, however, combined this method with the dimension reduction method known as PLS. Because PLS can be executed very efficiently and RLSC can also be speeded up using the orthogonal components generated from PLS as inputs, the new algorithm is computationally efficient and its performance is as good as or even better than SVM. We also used RFE to select a small subset of marker genes using this new algorithm and the results are very satisfactory.

Methods

Consider a microarray dataset containing n samples, with each sample represented by the expression levels of m genes. PLS is a technique for modeling a linear relationship between a set of input variables $\{\mathbf{x}_i\}_{i=1}^{n} \in R^m$ and a set of output variables $\{b_i\}_{i=1}^{n} \in R$. Only one-dimensinal output is considered here. So b_i = 1 or –1 corresponds to the i th sample belonging to class 1 or –1. Furthermore, we assume centered input and output variables. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \mathrm{K}\ , \mathbf{x}_n]^T$

and $\mathbf{b} = [b_1, b_2, \mathrm{K}\ , b_n]^T$. The PLS method finds the weight vectors **u** and **c** so that

$$\max_{|\mathbf{r}|=|\mathbf{s}|=1}[\mathrm{cov}(\mathbf{Xr}, \mathbf{bs})]^2 = [\mathrm{cov}(\mathbf{Xu}, \mathbf{bc})]^2 = [\mathrm{cov}(\mathbf{t}, \mathbf{d})]^2,$$

where **t** and **d** are score vectors for input vectors **X** and output vector **b**. After obtaining the pair of **t** and **d**, **X** and **b** are deflated by **t** and the procedure is repeated to obtain a new pair of **t** and **d**. The deflation rule described in the SIMPLS algorithm [3] is used in this paper. The score vectors **t** produced by SIMPLS are known to be mutually orthogonal. The score vector **t** (n x 1) and loading vector **u** (**m** x 1) can be obtained from

$$\mathbf{u} = \mathbf{X}^T\mathbf{b} \tag{1}$$

$$\mathbf{t} = \mathbf{Xu}, \|\mathbf{t}\| \to 1 \tag{2}$$

t is also called the PLS component. Then the deflation rules are given by

$$\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}(\mathbf{t}^T\mathbf{X}) \tag{3}$$

$$\mathbf{b} \leftarrow \mathbf{b} - \mathbf{t}(\mathbf{t}^T\mathbf{b}). \tag{4}$$

After deflation, a new pair of **u** and **t** can be generated using (1) and (2). This process can be iterated for p times so that a sequence of $\mathbf{t}_1, \mathbf{t}_2, \mathrm{K}\ , \mathbf{t}_p$ and $\mathbf{u}_1, \mathbf{u}_2, \mathrm{K}\ , \mathbf{u}_p$ are generated, assuming p is the number of PLS components required.

Now we formulate the RLSC classification method using the components produced by PLS. Let **T** be a (n x 1) matrix of which the p columns are PLS components such that $\mathbf{T} = [\mathbf{t}_1, \mathbf{t}_2, \mathrm{K}\ , \mathbf{t}_p]$. Also, let **w** be a ($n$ x 1) vector of weight components and $\mathbf{w}_0$ a (n x 1) vector with each element equal to the bias $\mathbf{w}_0$. The RLSC method solves the following optimization problem:

$$J(\mathbf{w}, w_0) = \|\mathbf{Tw} + \mathbf{w}_0 - \mathbf{b}\|^2 + C\|\mathbf{w}\|^2,$$

where C is the regularization parameter. Comparing this objective function with that of SVM, the squared loss function replaces the hinge loss function and all samples instead of support vectors are included in forming the classifier. Therefore, solving RLSC requires only a single system of linear equations, whereas solving SVM requires quadratic programming. In the case of microarray data, because the number of tissue samples is usually small (tens or hundreds) while the number of gene expression levels is very large (thousands or tens of thousands), solving these linear equations can be more efficient than quadratic programming. Now let $\mathbf{w}_1^T = [\mathbf{w}^T, w_0]$, $\mathbf{w}_2^T = [\mathbf{w}^T, 0]$ and $\hat{\mathbf{T}} = [\mathbf{T}, \mathbf{1}_n]$, where $\mathbf{1}_n$ is (n x 1)a vector with all elements equal one. The objective function can equivalently be written as

$$J(\mathbf{w}_1) = \left\|\hat{\mathbf{T}}\mathbf{w}_1 - \mathbf{b}\right\|^2 + C\|\mathbf{w}_2\|^2$$

To minimize J, we have

$$\frac{\partial J}{\partial w_1} = 2\hat{T}^T(\hat{T}w_1 - b) + 2Cw_2 = 0 \tag{5}$$

Remember that $\mathbf{T}^T\mathbf{T} = \mathbf{I}_p$ since the components produced by SIMPLS are normalized and mutually orthogonal and $\mathbf{I}_p$ is a (n x 1) identity matrix. Define $\sigma_i = \sum_j t_{ji}$, where t_{ji} is the i-th element of $\mathbf{t}_j$. From (5), $\mathbf{w}_1$ can be obtained from

$$\mathbf{w}_1 = \mathbf{R}^{-1}\hat{\mathbf{T}}^T\mathbf{b}, \tag{6}$$

where

$$\mathbf{R} = \begin{bmatrix} 1+C & & & \sigma_1 \\ & \ddots & & \vdots \\ & & 1+C & \sigma_p \\ \sigma_1 & \cdots & \sigma_p & n \end{bmatrix}. \tag{7}$$

It can be easily proved that $|\sigma_i| \le 1/2$. Thus $\mathbf{R}$ is always symmetric and positive definite and can then be conveniently inverted by menas of Cholesky decomposition [4], which requires less than half the time of a general decomposition. Furthermore, because the number p of components required by a classification problem is less than ten most of the time, the size of matrix $\mathbf{R}$ is relatively small and thus only very little computation is needed. In the experiments of this paper, p is empirically fixed to five until the number of features selected is less than ten, when we set p equal to the number of features. For gene selection purpose, we also want to know the weights for each gene. Thus, let $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_p]$, and the weight vector for the genes can be obtained by

$$\mathbf{v} = \mathbf{U}\mathbf{w}. \tag{8}$$

Denote $\mathbf{v}^* = |\mathbf{v}|$; the elements of $\mathbf{v}^*$ are the magnitudes of the weight vector $\mathbf{v}$ and they thus indicate the relative importance of the genes for classification. The RFE was used to select genes, which is described by Guyon *et al.* [5], and it is a procedure to select a series of nested subsets of genes by eliminating the genes that are least important for classification. For each subset of genes, we used leave-one-out cross-validation (LOOCV) to select the unknown parameter C. Then the bootstrap is used to generate 100 samples and calculate the averaged absolute weight vector $\overline{\mathbf{v}}^*$ for genes by training the classifier for each sample using the fixed C. $\overline{\mathbf{v}}^*$ is then sorted in descending order and the top ranked genes are selected. Eliminating an appropriate number of genes each time, this procedure is repeated until there is only one gene left. The performance of the classifier on each of these subsets can be assessed as follows: 1. The numbers of training samples and testing samples are firstly fixed and then 100 random partitions are performed; 2. For each partition, the unknown parameter C is determined by 10-fold cross-validation on training samples and then the classifier is built on the training dataset and tested on the testing dataset. Here the minimum 10-fold cross-validation error is called the training error. The training and testing errors are recorded and their means and standard deviations are calculated.

Results and Discussion

The colon cancer data published by Alon *et al.* [1] were used in the experiments. The dataset consists of 40 cancer tissues and 22 normal tissues with 2000 gene expression levels per sample. We first compared the performance of PLS-RLSC and SVM [7] on the colon cancer dataset without feature selection using the random partition method. The original dataset is randomly separated into a 40-sample training dataset and a 22-sample testing dataset for a hundred times. The results of mean training and testing errors (μ) with standard deviations (σ) and execution time are listed in Table 1. PLS-RLSC performed better than SVM and it required significantly less time. The performance gain can be explained by the dimension reduction as the preprocessing step, which avoids the possible data over fitting and eliminates noises in microarray data. The RFE procedure was then employed with PLS-RLSC on the colon cancer dataset. After recursively eliminating irrelevant genes until there is only one gene left, we have chosen a six-gene subset to achieve the minimum mean testing errors. The whole process requires less than 22 minutes. By using this subset, the mean training errors is known to be 0.59 ($\sigma = 0.67$) and the mean testing errors is known to be 0.47 ($\sigma = 0.73$). So the expected accuracy of the classifier on colon cancer data using this subset is estimated to be (22–0.47)/22 = 98 %. We listed the accession numbers and descriptions of the six genes selected in Table 2. Among the six genes, five (T57882, R88740, H08393, Z50753, H64807) are the same as those obtained in Shevade and Keerthi [8], who had chosen an eight-gene subset for colon cancer classification using sparse logistic regression.

Table 1. PLS-RLSC vs. SVM on the colon cancer data without feature selection

Method	Training errors (μ,σ)	Testing errors (μ,σ)	Time (s)
PLS-RLSC	6.10, 2.30	3.39, 1.16	243
SVM	6.95, 2.52	3.95, 1.49	1612

Table 2. Selected top six genes with their weights for the colon cancer data

Accession number	Description	Weight $\overline{\mathbf{v}}^{*}$)
T57882	Myosin heavy chain, nonmuscle type A (Homo sapiens)	0.50
R88740	ATP Synthase coupling factor 6, mitochondrial precursor (human)	0.50
X53586	Human mRNA for integrin alpha 6	0.41
H08393	Collagen alpha 2(XI) chain (Homo sapiens)	0.37
Z50753	H. sapiens mRNA for GCAP-II/uroguanylin precursor	0.27
H64807	Placental folate transporter (Homo sapiens)	0.26

References

1. Alon U., Barkai N., Notterman D.A., Gish K., Ybarra S., Mack D., Levine A.J. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays // Proc. Natl Acad. Sci. USA. 1999. V. 96. P. 6745–6750.
2. Brown M., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares Jr. M., Haussler D. Knowledge based analysis of microarray gene expression data using support vector machines // Proc. Natl Acad. Sci. USA. 2000. V. 97. P. 262–267.
3. de Jong S. SIMPLS: an alternative approach to partial least squares regression // Chemometrics and Intelligent Laboratory Systems. 1993. V. 18. P. 251–263.
4. Golub G.H., Van Loan C.F. Matrix Computations. The Johns Hopkins University Press, 1996.
5. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines // Maching Learning. 2002. V. 46. P. 389–422.
6. Rifkin R., Yeo G., Poggio T. Regularized least-squares classification, technical report. Available: http://www.ai.mit.edu/projects/cbcl/publications/ps/rlsc.pdf
7. Schwaighofer A. 2001. Available: http://www.cis.tugraz.at/igi/aschwaig/svm_v251.tar.gz
8. Shevade S.K., Keerthi S.S. A simple and efficient algorithm for gene selection using sparse logistic regression // Bioinformatics. 2003. V. 19, N 17. P. 2246–2253.

ON THE STATIONARY POINTS OF REGULATORY CONTOURS OF GENE NETWORKS

*Likhoshvai V.A.**[1,2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Contact author: e-mail: likho@bionet.nsc.ru

Keywords: *genetic systems, regulatory contours of gene networks, modeling, discrete methods, stable points, stationary points*

Resume

Motivation: Prediction of regimes of molecular-genetic system functioning by structural and functional organization of a system is one of the key problems in the fields of bioinformatics studying gene network functioning. To address this problems, it is necessary to perform theoretical studies of functioning of gene networks' regulatory contours and to reveal their general regularities, which determine the presence or absence of ability to support stationary, cyclic, or other, more complex regimes of functioning.

Results: In this work, we introduce genetic automates, as mathematical models of regulatory contours of gene networks. In terms of the graph theory, we give a description of the structure of stationary points of regulatory contours with the negative feedback.

Introduction

Gene networks (GN) are structurally complex spatial objects compiled by hundreds of elements of different nature and complexity: genes, their regulatory units, RNA and proteins encoded by genes; low-molecular substances, various complexes of enzymes and their targets, etc. (Kolchanov *et al*., 2000). GN elements are coupled into unique functional system by complex non-linear biochemical processes of synthesis and degradation of substances (Kolchanov *et al*., 2002). GNs are open systems, with their functioning supported by the constant influx into the media of particular substances and energy, as well as by efflux of metabolic products. GN functioning may be characterized by temporal trajectories illustrating alteration of concentrations of some integrity of substances circulating in GN.

Among the most important properties of GN is an ability to alter the state (substance concentration) in response to alteration of conditions of inner and outer environment. State alteration is produced via changing the level of particular gene group expression by substances-regulators. Regulatory processes are represented by the consequence of molecular events (frequently rather complex and branched), which may simultaneously involve regulatory gene regions and numerous substances, both coming from outside (external signals) and synthesized by the gene network itself (internal signals). So, the nucleus of GN is comprised by the genes and proteins encoded by genes, expression of which is mutually regulated. These sub-networks are regulatory contours of GN. Studying of the properties of regulatory contours is the pivotal task of bioinformatics dealing with the study of GN functioning, since exactly these properties supply GN by a unique ability to response adequately on alteration of external conditions.

In this work, we develop a discrete approach enabling to extract information about the gene network's stationaries directly from analysis of oriented graphs, which are represented by regulatory gene networks contours, thus, avoiding the stage of constructing and calculating dynamic models. The data obtained serve as an additional source of hypotheses on the properties of relevant dynamic models developed in (Likhoshvai *et al*., 2001, 2003, 2004; Fadeev, Likhoshvai, 2003).

Results

Let us set apart a section of gene network comprised of genetic elements that regulate expression efficacy of some genes or that are being controlled by other genes themselves. Let us denote this section as a regulatory contour of gene network. Then we represent the scheme of regulation of gene expression activity as a bipartite oriented graph $G(U_1, U_2, W_{12}, W_{21})$, with the vertices of the first type (U_1) identified with proteins; while the vertices of the second type (U_2), with independent mechanisms of regulation. By the arcs starting from W_{12} and passing from the vertices of the first type into the vertices of the second type, we denote entering of a protein into relevant mechanism of regulation, whereas arcs from W_{21} mean the action of the mechanism onto particular protein. On the vertices of the first type of bipartite graph $G(U_1, U_2, W_{12}\ W_{21})$ we construct the unipartite oriented graph $G(U_1, W)$, in which the arcs represent the entering of a particular protein into the mechanism of gene expression regulation, with the arcs directed to the proteins. A unipartite graph $G(U_1, W)$ we name as associated with the graph $G(U_1, U_2, W_{12}, W_{21})$.

Let us relate with each node of the first type u_i a non-negative variable x_i ranging within the limits $[0, p_i]$, where p_i is the upper limit of ranging of the i-th variable.

We order a non-negative number q, non-negative integer k and positive $s_1,\ldots,s_k$. By $B_0(y_1,\ldots,y_k|q, s_1,\ldots, s_k)$ we denote a Boolean function of k variables determined in $R_+^k = \{(y_1,...,y_k) \mid y_i \geq 0, i = \overline{1,k}\}$, which takes the value 0 then and only then $y_1^{s_1} \cdot \ldots \cdot y_k^{s_k} \leq q^{\sum_{i=\overline{1,k}} s_i}$. Otherwise, $B_m(y_1,\ldots,y_k|q, s_1,\ldots, s_k)=1$. If $k=0$, then by definition we set $B_0(\varnothing|q,0)= 0$. For negation $\neg B_0(y_1,\ldots,y_k|q, s_1,\ldots, s_k)$ we introduce the denotation $B_1(y_1,\ldots,y_k|q, s_1,\ldots, s_k)$. The simplest biological prototypes of the functions introduced are: for B_0, the mechanism of the threshold inhibition (for B_1, the mechanism of the threshold activation) of transcriptional activity of genetic element by a multimer compiled by s_1 molecules of the form $y_1,\ldots, s_k$ molecules of the type y_k. In general case, $s_1,\ldots, s_k$ are real numbers called Hill's coefficients, which order the degree of nonlinearity of participation of respective effector in a regulatory mechanism.

Let i be a number of the node of the first type. The set of numbers of vertices, out of which the arcs of the digraph $G\ (U_1, W)$ proceed to u_i, we denote via D_i. Let l_i arcs are directed towards the i-th node of the second type. Let them be numbered as $\sigma_i(j)$, $j=1,\ldots, l_i$. By $R_{i,j}$, we denote a set of numbers of vertices of the first type, with the arcs directed to the node of the second type enumerated by $\sigma_i(j\ )$. If the node of the second type is rooted, then by definition, we expect $R_{i,j}=\varnothing$. Let us relate to each node of the second type the integer $\delta_{i,j} \in \{0,1\}$ and function $B_{\delta_{i,j}}$.

Then we compose a Boolean function

$$B_i(x_j \big| j \in D_i) = \underset{j:\delta_{i,j}=1}{\vee} B_{\delta_{i,j}}(x_r \big| r \in R_{i,j}, q_{i,j}, s_{i,j,1},..., s_{i,j,l_{i,j}}) \vee (\underset{j:\delta_{i,j}=1}{\wedge} B_{\delta_{i,j}}(x_r \big| r \in R_{i,j}, q_{i,j}),$$

which describes integral mechanism of regulation of activity of the i-th genetic element.

Definition of genetic automate

Let us associate with each node of the first type three non-negative numbers, α_i, δ_i, p_i and the variable x_i taking on even values in the interval $[0,p_i]$. We denote a function $G_p(x_1,...,x_n) = (x_1^+,...,x_n^+)$, where

$$x_i^+ = \begin{cases} \max(\ 0, x_i - \delta_i)\ ,\ B_i(x_j \big| j \in D_j) = 1 \\ \min(\ p_i, x_i + \alpha_i)\ ,\ B_i(x_j \big| j \in D_j) = 0 \end{cases},$$

a genetic automate (*G*-automate).

The consequence of points

$$X^0 = (x_1^0,...,x_n^0), X^1 = (x_1^1,...,x_n^1) = G(x_1^0,...,x_n^0), X^2 = G^2(x_1^0,...,x_n^0),...\ X^k = G^k(x_1^0,...,x_n^0),\ ...$$

is named a trajectory of the *G*-automate (starting from the point X^0). One action of an automate is called a tact. Obviously, due to finiteness of the space of vectors-values, any trajectory of genetic automate after finite number of tacts becomes cyclic. By cycle length, we denote the minimal number of non-recurring points of a trajectory. If the cycle length equals to 1, then it is called a motionless point.

We set the task to describe the stationary points of genetic automates. Biological meaning of this description is that the stationary points of genetic automates correspond to the stationaries of genetic networks, which have respective structures of regulatory contours.

Considerations given below give evidence that the stationary points do not depend upon the values of an automate's parameters, that is, they are determined only by the structure of bipartite digraph. Really, let us admit that genetic automate with some ordered set of parameters $q_{i,j}, s_{i,j,1},...,s_{i,j,l_{i,j}}$ α_i, δ_i, p_i has a stationary point. Let us take some non-zero threshold (a parameter indicated by *q* with an index) and decrease it for some minor value. Obviously, the stationary point will stay the same. By decreasing the threshold value smoothly, we may bring it to zero. By analogy, we may consequently bring to zero all the rest thresholds of genetic automate. Similarly, we may bring all the parameters of the groups α, δ, and *p* to the value 1. After this, we note that stoichiometric coefficients (s) also could be set as equaling to 1. Therefore, for solving the problem of detecting stationary points, we may limit ourselves by considering two-digit genetic automates with zero thresholds.

***Definition**. An integrity V of the first type vertices is a g-base (of the bipartite, associated digraph) then and only then: (i) if the node u_i is a root node of associated digraph, then it obligatory lies in V, (ii) if the node u_i from V is not a root node of associated digraph, then 1) for any non-empty $R_{i,j}$, such that δ$_{i,j}$ =0, ∃ the node with number from U_1\V, 2) either there does not exist $R_{i,j}$, such that $\delta_{i,j}$ =1 or always there exists at least one $R_{i,j}$, for which δ$_{i,j}$ =1, such that any node with number from $R_{i,j}$ belongs to V, (iii) for any node u_i from U_1\V, either ∃ non-empty $R_{i,j}$, for which δ$_{i,j}$ =1, such that for some node with number from $R_{i,j}$ belongs either U_1\V, or ∃ non empty $R_{i,j}$, for which δ$_{i,j}$ =0, such that all vertices with numbers from $R_{i,j}$ belong to V.*

Biological meaning of this definition is simple. All vertices from V correspond to active genes, while vertices from the complement, to passive ones. Hence, all constituently expressed genes should enter the variety of active genes, this evidence being ordered by condition (i). On the other hand, in order the regulated genetic element be expressed, it should be activated or should have non-zero basal level of activity. In order some mechanism of activation be functional, all activators composing this mechanism should be expressed, that is, condition (ii2) should be valid, as well all inhibitory mechanisms should be switched off, hence, condition (ii1) assert. On the contrary, all genetic elements, respective vertices of which belong to U_1*V* will be passive only under performance of condition (iii). So, it occurs that this definition describes all stable points of genetic automates.

Lemma. Let us have constructed a genetic automate on the bipartite digraph *G(*U_1, U_2, W_{12}, W_{21}*)*. Then any *g*-base *V* in *G(*U_1, U_2, W_{12}, W_{21}*)* generates in genetic automate the stationary point of the form x_i=1, if $x_i \in V$, otherwise, x_i=0. The opposite assertion is also valid. For any rest point of genetic automate constructed on pre-ordered bipartite digraph $G(U_1, U_2, W_{12}, W_{21})$, the sub-set of vertices of the first type, with non-zero values of variables x_i, is a g-base.

Discussion

The problem of studying general regularities of natural gene networks functioning is extremely complicate due to unique gene network composition. Studying of hypothetical constructions designed in accordance with some pre-ordered rules may alleviate the solution of this problem. For example, it is possible to stand out genetic elements and regulatory mechanisms as the constructional elements and then to construct out of them a hypothetical gene network [3–6]. By assuming that regulatory mechanisms are composed by some integrity of independent events, which consist of the threshold-like interactions between complexes with sites-targets, we arrive at description of gene networks' regulatory contours by genetic automates. The result obtained demonstrates that out of supposition on threshold-like mechanism of regulatory mechanism action, it follows that existence of stationaries is completely determined by structural and functional organization of the gene networks' contours represented by bipartite digraph $G(U_1, U_2, W_{12}, W_{21})$ and undependable upon the model's parameters. The resulted description of stationaries of genetic automates in terms of g-bases gives theoretical solution of the problem of determining stationary points of arbitrary gene network regulatory contours. To this aim, it is necessary to construct the relevant bipartite digraph and to find all its g-bases. In parallel, we detect the structure of every stationary point. The practical implementation of this result demands to develop an algorithm searching for g-bases and to design the appropriate software, which is a technical task.

The result obtained poses the important theoretical problem of searching for right criterions for detection of g-bases. In general case, this problem is a generalization of the problem of description of 1-base of an oriented graph (for definition of 1-base, see Harary, 1969). Really, if we limit ourselves by considering regulatory contours with regulatory mechanisms controlled by homomultimers, referring to the negative type of action, then g-bases are identical to 1-bases. Since the problem of describing 1-bases is actual for more than 25 years and still is not solved, it is unlikely to expect that in general case, simple criteria for searching for g-bases will be developed. More probably, such criteria could be constructed for some classes of digraphs. For example, to begin with, it is of interest to describe g-bases for the simplest class of digraphs, which admit a rotational group ordered by permutation (12…n), where n is the number of vertices in a digraph.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (Nos. 02-04-48802, 02-07-90359, 03-01-00328, 03-04-48506, 03-04-48829, 03-07-96833, 04-01-00458), Russian Ministry of Industry, Sciences and Technologies (No. 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Project No. 119), and RAS Presidium Program "Molecular and Cellular Biology" (Project No. 10.4).

References

Fadeev S.I., Likhoshvai V.A. On hypothetical gene networks // Sib. J. of Industrial Mathematics. 2003. V. 6(15). P. 134–153. (In Russian).

Harary F. Graph Theory. AddisonWesley, Reading, MA, 1969. 274 p.

Kolchanov N.A., Ananko E.A., Kolpakov F.A., Podkolodnaya O.A., Ignatieva E.V., Goryachkovskaya T.N., Stepanenko I.L. Gene networks // Mol. Biol. 2000. V. 34(4). P. 533–544. (In Russian).

Kolchanov N.A., Ananko E.A., Likhoshvai V.A., Podkolodnaya O.A., Ignatieva E.V., Ratushny A.V., Matushkin Yu.G. Gene networks description and modeling in the GeneNet system, Chapter 7 // Gene Regulation and Metabolism / Eds. J. Collado-Vides, R. Hofestadt. The MIT Press, Cambridge, Massachusetts. 2002. P. 149–180.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. Relationship between a gene network graph and qualitative modes of Its functioning // Mol. Biol. 2001. V. 35(6). P. 926–932.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. The problems of theory of gene networks functioning // J. of Industrial mathematics. 2003. V. 6. P. 64–80. (In Russian).

Likhoshvai V.A., Fadeev S.I., Matushkin Yu.G. The global operation modes of gene networks determined by the structure of negative feedbacks // Bioinformatics of genome regulation and structure / Ed. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London. 2004. P. 319–330.

MODELLING OF SUBSTANCE SYNTHESIS PROCESS WITHOUT BRANCHING BY THE DELAY EQUATION

Likhoshvai V.A.[1,3], **Demidenko G.V.**[2], **Fadeev S.I.**[2]*

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Sobolev Institute for Mathematics SB RAS, Novosibirsk, Russia; [3] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia

* Corresponding author: e-mail: likho@bionet.nsc.ru

Keywords: *genetic systems, modeling, delay equation, differential autonomous systems*

Summary

Motivation: Specificity of a gene network as an object of mathematical and computer modeling is determined by the fact that it refers to the class of super-large systems, with flows of substances and energy expressed via the synthesis of many dozens and hundreds of thousands intermediate forms of DNA, RNA, and proteins. Accounting for intermediate stages of synthesis of DNA, RNA, and proteins emerges the systems of ordinary differential equations with huge dimensionality. Hence, a necessity appears to develop theoretical and numerical studies aimed at searching for conditions that favor to diminishing of dimensionality of the gene networks models without loosing their adequacy.

Results: In the work presented, we consider a substance synthesis model without branching. The theorem on the limited transition from mathematical model in a form of autonomous system of equations to the mathematical model described by delay equation under tending to infinity of the number of equations describing intermediate stages of synthesis.

Introduction

Studying of regularities of gene networks functioning is one of pivotal problems of post-genome molecular biology and genetics. With this aim in view, the methods of computer and mathematical modeling are widely used (Likhoshvai *et al.*, 2003; Edwards, Glass, 2000; Elowitz, Leibler, 2000; Gardner *et al.*, 2000; Thomas *et al.*, 1995). Specifics of gene networks as an object of mathematical and computer study lies in the fact that they belong to super-large systems, in which the flows of substance and energy are mediated by synthesis of many dozens and hundreds of thousands intermediate forms of DNA, RNA, and proteins. The synthesis of these substances is executed by fundamental and multi-stage processes of replication, transcription, and translation. Accounting for intermediate stages of DNA, RNA, and protein synthesis emerges the differential equations systems with huge dimensionality. So, a demand arises to decrease dimensionality of gene network models without the loss of their adequacy. One of the approaches that could be used in modeling is dividing the processes into rapid and slow with subsequent reduction of the differential equation systems on the basis of Tikhonov's theorem (Tikhonov, 1952). The second approach is based on elimination from the model of some variables due to these or that considerations that emerge from specificity of the modeled system and/or from the essence of the task to be solved. As a rule, these considerations are poorly described by exact methods and are based on semi-intuitive considerations. For example, under modeling, a group of subsequent processes is frequently substituted by delay parameter that equals numerically to the summarized duration of the processes. In the study presented, we give the exact grounding of adequacy of this approach application for reducing the model in the class of substance synthesis without branching. This class of models frequently appears as the constituent part of more general gene networks models, because they help to describe the processes of replication, transcription, translation, and enzyme chain reactions. In this work, we prove the theorem on the limiting transition from the system of autonomous equations to differential system with delay, under tending to infinity of the number of equations describing the intermediate stages of synthesis. We have also constructed the functions and biases appearing during transition from the finite models to delay equation.

Results

We consider a mathematical model of irreversible multi-stage substance synthesis process without branching:

$$\begin{cases} dy_1/dt = f(y_n) - (n-1)\tau^{-1} y_1 \\ dy_i/dt = (n-1)\tau^{-1}(y_{i-1} - y_i) \ , \ i = \overline{1, n-1} \\ dy_n/dt = (n-1)\tau^{-1} y_{n-1} - \theta \ y_n \ . \end{cases} \tag{1}$$

where y_i is a concentration of intermediate stages of the protein synthesis, y_n, concentration of the final product of synthesis; $f(y_n)$, rather smooth function describing the regulatory mechanism of initiation of the synthesis of a substance; τ is the total time needed to proceed via the stages from the 1-st condition to the n-th condition; θ, the constant of dissipation rate of the final product from reaction mixture.

From (1), we arrive at

$$y_n(t) = \int_o^t \psi_n(t-s)\ f(y_n(s))\,\mathrm{ds},\ \psi_n(t) = e^{-\theta t} / (1 - \tau\theta(n-1)^{-1})^{n-1}\ S_n(t), \tag{2}$$

where $S_n(t) = 1 - e^{-rt} \sum_{k=0}^{n-2} (r\,t)^k / k!$, $r = (n-1)\tau^{-1} - \theta$. In (2), we observe the limiting transition $\lim_{n\to\infty} y_n(t) = x(t)$, $\lim_{n\to\infty} \psi_n(t) = \begin{cases} 0, & t < \tau, \\ \exp(\theta\ (\tau - 1)), & t > \tau \end{cases}$, where x(t) satisfies to equation with delay

$$dx/dt \ = f(x(t-\tau)) - \ \theta\ x\ , \quad t > \tau\ .$$

As follows from two theorems given below, the function $S_n(t)$ converges evenly to the Heaviside's unit function under n→∞ for the interval [0,τ-ε], [τ+ε,T] by t for every ε > 0 and, therefore, the limited representation is $\psi_n(t)$.

Theorem 1. Let $t = p\tau$, $p > 1$, $n_p = [p\theta\tau/(p-1)] + 1$. Then under $n \geq n_p$, we arrive at the estimate:

$$| S_{n+1}(t) - 1 | < \left(\sqrt{2\pi(n-1)}(p(1-\theta\tau/n)-1)\right)^{-1} (\,pe^{1-p}\,)^n e^{\theta t}(1 - \theta\tau/n\,)^n.$$

Theorem 2. Let $t = \tau/p$, $p > 1$. Then for $n > \tau\theta$, we arrive at the estimate: $| S_{n+1}(t) | <$ $\left(\sqrt{2\pi n}(1-(1-\theta\tau/n)/p)\right)^{-1} (\,e^{1-(1-\theta\tau/n)/p}(1-\theta\tau/n\,)/p)^n$.

On the basis of the theorems 1-2, we prove that the function $y_n(t)$ converges uniformly to the function $x(t)$, $n \to \infty$, within some interval $[0, T]$ and evaluate deviation as $\max_{t\in[0,T]} |y_n(t) - x(t)|$, $n >> 1$.

Theorem 3. Assume that the function $g(z)$ satisfies to Lipschitz condition:

$$|f(z_1)-f(z_2)| \le L|z_1-z_2|, \qquad z_1, z_2 \in R.$$

Let $T > \tau$, such that $L(1-e^{-\theta\tau})/\theta < 1$. Then the sequence $\{y_n(t)\}$ converges uniformly within the interval $[0,T]$.

Theorem 4. Let $x(t) \in C[0,T]$ be the limiting function of the sequence $\{y_n(t)\}$. Then there exists an estimate $\max\limits_{t\in[0,T]}|y_n(t)-x(t)| \le \left(1-L(1-e^{-\theta T})/\theta\right)^{-1} I_n$, $n >> 1$, where $I_n = G\left(A_n(1-e^{-\theta T})/\theta + \quad n^{-1/4}(1-\theta\tau/(n-1))^{1-n}\left(3(1-e^{-\theta T})/\theta+8\tau\right)\right)$, $A_n = e^{\theta\tau} - (1-\theta\tau/(n-1))^{1-n}$.

As obviously follows from definition I_n, we have $I_n \to 0$, $n \to \infty$. Hence inequality indicated in the theorem 4 gives uniform estimate in the interval $[0,T]$ of the *n*-th approximation to the limiting function *x*(*t*), which is the solution of the integral equation (4) due to the theorems 1–3.

Discussion

The systems of equations of the form (1) appear as the constituent elements of more general differential equations systems that model gene networks, because network functioning is based on such fundamental matrix processes as replication, transcription and translation, which could be referred in a first approximation to irreversible processes composed from large number of consequent rapidly processed intermediate stages. The entry of y_n in the right part of the first equation of the system (1) appears if there exists a regulation (repression, activation) of the effectiveness of a process by its final substance (product). From this viewpoint, studying of the properties of the system of the form (1) and its generalizations is an important task of the theory of gene network modeling.

The result obtained gives evidence that if the synthesis has sufficiently large number of linear stages, and the rate of processing of each intermediate stage is rather high, then kinetics of production of the final product is almost undependable kinetics of inner stages of synthesis. The whole process is determined by the mechanism of regulation of the synthesis initiation (launching the first stage of synthesis) and the value of delay, which equals to the average summarized time of duration of all intermediate stages. In other words, the result obtained in this work estimates the relationships between the micro- and macro-levels of the system's functioning, in case we consider stages of synthesis for a micro-level and the final product, for a macro-level, respectively. This relationships may be expressed as the following statement: *A single stage of synthesis occurring at the micro-level is less influences kinetics of the final product production if the lesser is the time it occupies in the whole integrity of subsequently occurring micro-processes.* In the limit at the macro-level, only one characteristics of micro-level is revealed, namely, the summarized time of duration of the process of synthesis.

With respect to suggested interpretation of the result obtained, a natural questions arise on criticality of linearity conditions and reversibility of intimidate stages, which are necessary for validation of the limited theorem proved. Indeed, in real biological systems, separate stages of DNA, RNA, and protein synthesis are linear and irreversible only in the first approximation. In general, they are non-linear, because they are represented by integrity of biochemical reactions. Due to the same reasoning, the stages of synthesis lose irreversibility to more and more extent when we disintegrate them, gradually approximating to the level of elementary biochemical events.

So, the studying of limiting transitions in the systems describing multi-stage processes under different mechanisms of intermediate stages of synthesis is very important for construction of gene networks' theory. Justification of the limiting transition makes a theoretical basis for intuitive understanding of the fact that for adequate modeling of processes at the macro-level, the knowledge on gene network functioning at every micro-level stage is not necessary. In future, we plan to develop the limiting theory towards attenuating conditions set for the system (1).

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (Nos. 02-04-48802, 02-07-90359, 03-01-00328, 03-04-48506, 03-04-48829, 03-07-96833, 04-01-00458), Russian Ministry of Industry, Sciences and Technologies (No. 4 3.073.1.1.1501), Siberian Branch of Russian Academy of Sciences Integration Projects Nos. 119, 148, and RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4).

References

Elowitz M.B., Leibler S. A synthetic oscillatory network of transcriptional regulators // Nature. 2000. V. 403. P. 335–338.

Gardner T.S., Cantor C.R., Collins J.J. Construction of a genetic toggle switch in *Escherichia coli* // Nature. 2000. V. 403. P. 339–342.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. Problems of the theory of gene networks functioning // Sib. J. of Industrial mathematics. 2003. V. 4. P. 64–80. (In Russian).

Thomas R., Thieffry D., Kaufman M. Dynamycal behavior of biological regulatory networks-I. Biological role of feedback loops and practical use of the concept of the loop-characteristic state // Bulletin of Mathematical Biol. 1995. V. 57. P. 247–276.

Tikhonov A.N. Systems of differential equations containing small parameters // Math. Sbornik. (Mosc.). 1952. V. 31. P. 575. (In Russian).

GENE NETWORK OF THE ARABIDOPSIS DEVELOPING SHOOT MERISTEM AND ITS DESCRIPTION IN THE GENENET COMPUTER SYSTEM

Mironova V.V., Omelianchuk N.A.*

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

Keywords: *SAM development, gene network, Arabidopsis*

Summary

Motivation: The so far accumulated data on gene regulation in the Arabidopsis developing shoot apical meristem (SAM) allow integrating the data within a gene network for their further analysis and modeling.

Results: Data on the regulation of gene expression were input into the GeneNet computer system. The gene network consists of indirect interaction in cells undergoing differentiation from the stem cells into those of the peripheral zone and later into the leaf primordium cells, and also of a spatial feedback loop that maintains the central zone size constant. The two distinguishing features of the gene network of the developing SAM are: (1) all or a part of the descendants of a differentiating cells are the founders of the next compartment and (2) a certain number of undifferentiated cells are retained by the feedback loop, involving both negative and positive interactions. Distinct location of constituent regulatory proteins in the SAM is the condition necessary for this feedback loop to function.

Availability: http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/

Introduction

All the cells of the aerial part of the mature plant are descendants of the stem cells of the SAM, which are at the growing tip. The SAM is divided into 3 zones. The central zone (CZ) includes the entire reservoir of the plants stem cells, whose descendants are displaced outward to the peripheral and rib zones. The cells of the peripheral zone (PZ) differentiate into the leaf primordium. The SAM is also divided into layers: the epidermal (L1), the subepidermal layer (L2), and the corpus. In turn, the surface layer and the organization center (OC) are distinguished in the corpus. The OC is required for the maintenance of the constancy of the content and size of the CZ. The study was aimed at the creation, on the basis of the published papers, of a formalized description for the gene network of the developing SAM in the GeneNet computer system. Our task was also to identify the group of genes controlling some of the SAM functions.

Methods and Algorithms

The data published in the papers were input into the GeneNet system (Ananko *et al.*, 2002). In the majority of publications concerning genes expressing in the Arabidopsis SAM, the expression of different genes in wild type, mutants and transgenic plants was compared. This allowed representing the gene network as indirect regulatory interactions. The references used for gene network construction are omitted because all are listed in the GeneNet database.

Results and Discussion

Maintenance of CZ size. A main mechanism maintaining the constancy of the CZ size is the spatial feedback loop, involving both negative and positive interactions (Fig.) between the OC and the CZ (reviewed in Sharma *et al.*, 2003). The key gene in the feedback loop is CLV3, which is normally transcribed only in the three CZ upper layers, with a few cells being transcribed in the third layer. The CLV3 protein enters the extracellular apoplast and migrates within CZ. When the signal for CZ is delivered as the CLV3 ligand, the RLK-LRR kinase CLV1 is autophosphorylated and forms the active CLV1/CLV2/KAPP/Rop complex. In the complex, KAPP may be modulated, while the Rop protein is involved in the further signal transduction through MAPK. As a result, the protein complex suppresses WUS expression in the surface layer of the corpus (the negative feedback loop). In the cells underlying the surface layer, the CLV1/CLV2/KAPP/Rop complex is not formed presumably because there is no CLV3 ligand completely consumed by the surface layer.

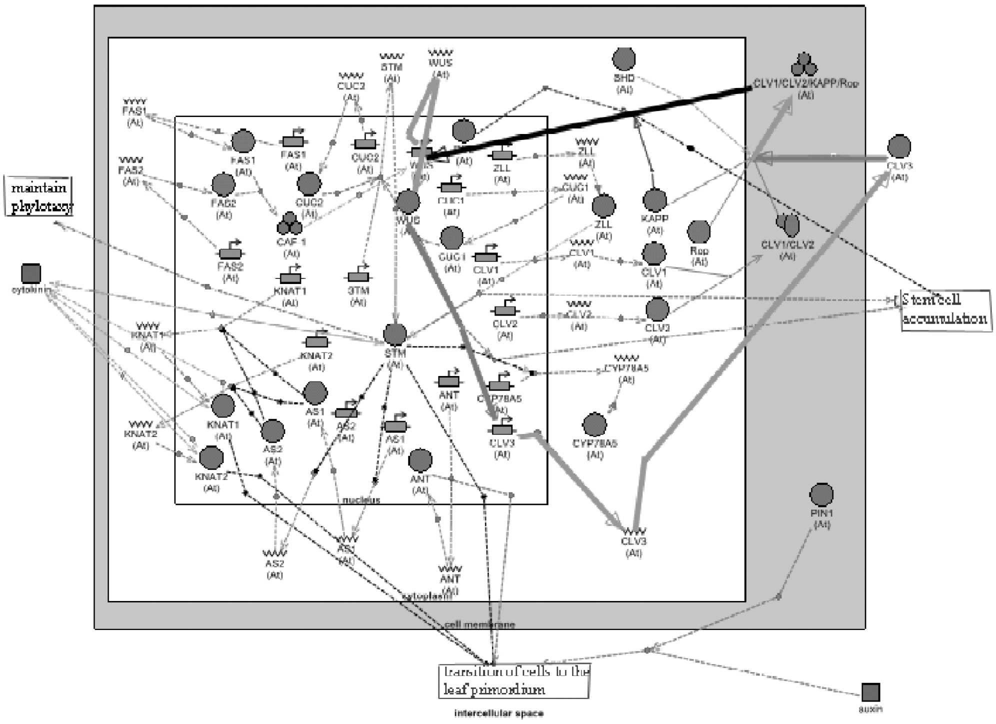


Fig. Gene network of the *Arabidopsis* developing SAM. Filled ovals denote the proteins; filled rectangles, genes; and arrows, regulatory effects. Bold lines display a spatial feedback loop, involving both positive and negative interactions.

These cells form the OC. Cell reference to the OC is entirely determined by the expression of the WUS gene in these cells (Mayer *et al.*, 1998). The WUS, by a so far unknown mechanism, increases CLV3 expression in the above lying CZ (the positive feedback loop). At a decreased WUS expression domain, CLV3 is less activated and as a result the CLV3 expression domain (the CZ) is reduced followed by an increase in the size of the WUS expression domain. And, in contrast, at an increase in the CZ size, the number of CLV3 increases in the intercellular space, this affects the decrease in the OC size resulting in a reduction in the WUS expression and thereby ultimately reducing the CZ size. Thus, there is a spatially disconnected feedback loop. Such a dynamic balance is disrupted by mutations in WUS and CLV3. Under the effect of the clv3 mutation, both the wus-expression domain and the CZ size increase. The primary SAM of the wus mutants is virtually devoid of the OC and CZ.

There is yet another independent mechanism designed to maintain the CZ size constant. The SAM-specific STM gene plays a major role in the process. Thus, in the strong stm mutant, there are no CZ and accordingly no stem cells and, instead partially differentiated cells expressing AS1 and AS2 proteins replace them (Byrne *et al*., 2002). It was shown that STM suppresses the expression of the AS1 and AS2 genes. At the time STM starts expressing in the embryo, the SAM begins to form. The CUC1 and CUC2 genes expressing at the early stages in the apical domain of the embryo induce the expression of the STM gene (Aida *et al*., 1999). Also, these genes specify the boundaries of the STM expression, thereby contributing to the SAM zonation in the mature embryo.

Transition of cells to the peripheral zone. The PZ is actually the first stage of the differentiation of the CZ cells. At this stage, CLV3 expression ends and the KNAT1, KNAT2 and ANT genes start to express. The PZ cells become sensitive to the action of hormones cytokinin and auxin.

Formation of the leaf primordium. The next differentiation stage of the stem cells is initiation of the leaf primordium. At stage 0 of the leaf primordium, there is no STM expression and, as a result, the AS1 and AS2 genes start to express, and they in turn suppress the expression KNAT1 and KNAT2 (Byrne *et al*., 2002).

Division of the gene network into compartments. The gene network of the Arabidopsis developing SAM is characterized by spatial disconnection of the stages of differentiating cells, i.e. cells at different stages form distinct compartments. There are two types of interaction between the compartments: regulatory relations and transition from one compartment to another as cells divide. In the CZ, we distinguish two compartments, the CZ tunica and the CZ surface layer of the corpus. The CZ tunica expresses CLV3 and shows no expression of the CLV1 and WUS genes. The CZ tunica cells divide in an anticlinal plane. We refer a few cells expressing CLV1 in the tunica L2 to the next compartment, the surface layer of the corpus. We suggest that these cells may be the descendants of the corpus cells, because the corpus cells divide randomly in all planes. Features of the corpus surface layer are expression of CLV3 and CLV1 and no expression of the WUS gene. In this layer, WUS is indirectly suppressed by the CLV1/CLV2/KAPP/Rop complex. The OC, the PZ and the leaf primordium are referred to the other compartments. Thus, the major feature of the gene network for Arabidopsis developing shoot meristem is dependency of its normal functioning and development on the proper formation of the compartments by the descendants of the differentiating stem cells. The required number of undifferentiated cells in the CZ is retained by the feedback loop, involving both positive and negative interactions. The condition for the correct function of the feedback loop is the spatial disconnection of its constituent regulatory proteins. Every compartment of the shoot apex shows differential gene activity.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (grants No. 02-07-90359, 03-07-96833, 03-04-48506-a, 03-07-90181-в, 03-04-48469-a), Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Projects No. 119), and US National Science Foundation FIBR PR 03-106 “Developing modeling and bioinformatics”, PAS Presidium Program “Molecular and Cellular Biology” (grant No. 10.4).

References

Aida M., Ishida T., Tasaka M. Shoot apical meristems and cotyledon formation during Arabidopsis embryogenesis: interaction among the CUP-SHAPED COTYLEDON and SHOOT MERISTEMLESS genes // Development. 1999. V. 126. P. 1563–1570.

Ananko E.A., Podkolodny N.L., Stepanenko I.L., Ignatieva E.V., Podkolodnaya O.A., Kolchanov N.A. GeneNet: a database on structure and functional organization of gene networks // Nucleic Acids Res. 2002. V. 30. P. 398–401.

Mayer K.F.X., Shoof H., Haecker A., Lenhard M., Jurgen G., Laux T. Role of WUSHEL in regulating stem cell fate in the *Arabidopsis* shoot meristem // Cell. 1998. V. 95. P. 805–815.

Moussian B., Schoof H., Haecker A., Jurgen G., Laux T. Role of the ZWILLE gene in the regulation of the central shoot meristems cell fate during Arabidopsis embryogenesis // EMBO J. 1998. V. 17. P. 1799–1809.

Sharma V.K., Carles C., Fletcher J.C. Maintaince of stem cell populations in plants // PNAS. 2003. V. 100. P. 11823–11829.

COMPUTER MODELING OF THE FUNCTION OF TRANSCRIPTION FACTORS DURING MACROPHAGE ACTIVATION

***Nedosekina E.A.*[1]*, *Ananko E.A.*[1], *Likhoshvai V.A.*[1, 2]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Corresponding author: e-mail: nzhenia@bionet.nsc.ru

Keywords: *mathematical simulation, model, gene networks, macrophage activation, transcription factors*

Summary

Motivation: Operation of the immune system is an intricate process comprising interaction and fine regulation of a large number of cells. So far, the molecular mechanisms underlying the overall operation of the immune system and functions of individual cell types yet remain vague. Computer modeling allows certain specific characteristics of the behavior of the gene network in question to be studied as well as the quantitative characteristics of this system to be determined.

Results: Specific features of the function of transcription factors during macrophage activation were considered using the previously developed computer model.

Introduction

Accurate and coordinated operation of the immune system underlies the normal function of the body, its protection from infection, and from penetration of the foreign objects. The essential part of the immune system is macrophages, synthesizing various cytokines, receptors, enzymes, and other compounds. Production of these substances increases drastically upon macrophage activation. It is known that the activated macrophage expresses over 200 genes, and in this case, each component of the gene network should be synthesized at proper time. Upon disappearance of the infection focus or stimuli, activation is disappearing and the concentration reverses to the norm. Timely response of the cell depends on the accurate work of the signal transduction pathways and the transcription factors. However, molecular basis of expression control during macrophage activation requires further studies.

Earlier, we developed a computer model of macrophage activation under the effect of lipopolysaccharides (LPS) and interferon-γ (IFN-γ; Nedosekina, 2002). The model was used to study the signal pathways of the gene network of macrophage activation inducing transcription factors (Nedosekina, 2003). In this work, we consider specific features of the functions of transcription factors controlling the gene network of macrophage activation under the effect of LPS.

Methods and Algorithms

A generalized chemical kinetic simulation method forms the basis of our computer model (Likhoshvai *et al.*, 2001). To construct the model, each elementary process (regulatory effect or reaction) was described using a standard block based on chemical kinetic equations. For example, formation of heteromeric protein comprising two different subunits may be described using the below reversible bimolecular reaction:

$$A + B \underset{k_2}{\overset{k_1}{\leftrightarrow}} C\ ;\quad \frac{dC}{dt} = -k_2 C + k_1 AB = -\frac{dA}{dt} = -\frac{dB}{dt}\ .$$

In the case the external stimuli are absent, this corresponds to the system of the ordinary differential equations with dynamic variables expressing concentrations of genes, mRNA, proteins, and low molecular weight substances. In general case, the mathematical model consists of differential equations and discrete expressions.

The key stage in construction of a model is verification of its parameters. For this purpose, published experimental data were used—initial approximation values of the model parameters were specified on the basis of this information. Further, the optimal values of these parameters searched for by original software (Likhoshvai *et al.*, 2002), which applies the method for simulation of the evolution of a population of individuals, were determined. In this method, it is supposed that each of individuals have individual set of values of the parameters, which are generated randomly. The most fit individuals are selected at each stage of the evolution. The fitness is considered as similarity of the results obtained by the model to the experimental data published.

Implementation and Results

Earlier, we developed a computer model of macrophage activation under the effect of LPS and IFN-γ (Nedosekina, 2002; *in silico*). Here, we are considering the specific features of operation of the transcription factors (TFs) under the effect of LPS.

Data of published experimental works on TF concentration dynamics were used to search for parameters of the mathematical model. For example, Fig. 1 shows the comparison of curves of changes in AP-1 TF concentration: experimental data (Hambleton, 1996) and results obtained using the model upon selection of the parameters using these data.

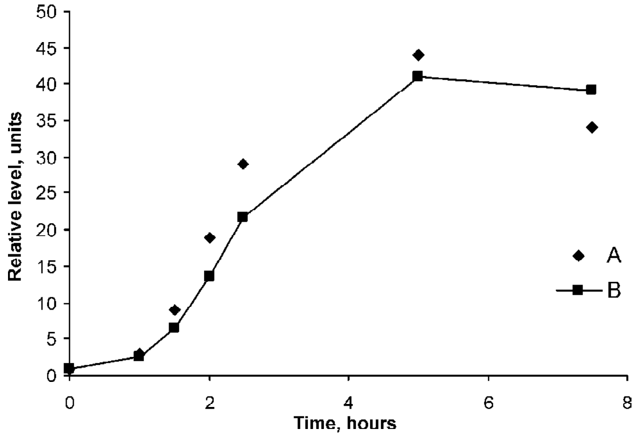


Fig. 1. AP-1 transcription factor concentration after LPS influence: A – experiment data (Herrero *et al.*, 2001); B – computer model.

However, deficiency in experimental data is the main problem in mathematical modeling. If the information on TF concentration in macrophage and its changes during activation was absent, the parameters were selected basing on indirect data on operation of signal transduction pathways or expression dynamics of the genes controlled by this TF.

For example, shown in Fig. 2A are the curves of changes in IRF-1 (interferon regulatory factor-1) mRNA and ICSBP (interferon consensus sequence binding protein, IRF-8) mRNA as well as concentrations of TFs controlling expression of these compounds—Stat-1α (Fig. 2B) and NF-κB (Fig. 2C). Data of experimental works used to construct the model demonstrate that expression of the genes *ICSBP* and *IRF-1* is regulated by Stat-1α and NF-κB transcription factors. The results of computations using the model demonstrate that the concentration of ICSBP mRNA increases 13-fold; IRF-1 mRNA, regulated by the same TFs, increases 42-fold (Fig. 2A). This difference may be reached due to specific features of binding sites for these transcription factors and certain additional proteins.

Shown in Fig. 3 are the results of calculations—curves of concentration dynamics of the following transcription factors upon the effect of LPS (0.01 μg/ml): AP-1, IRF-1, ICSBP, NFκB, p38 MAPK, and Elk-1. The curves are constructed in arbitrary units for each TF. As is evident from Fig. 3, concentrations of the TFs increase severalfold ICSBP, NFκB, Elk-1) and several dozen-fold

(AP-1, IRF-1, p38 MAPK), whereas Stat-1α TF concentration increases by several orders of magnitude (neither conformation nor disproof of this fact was found in the published data).

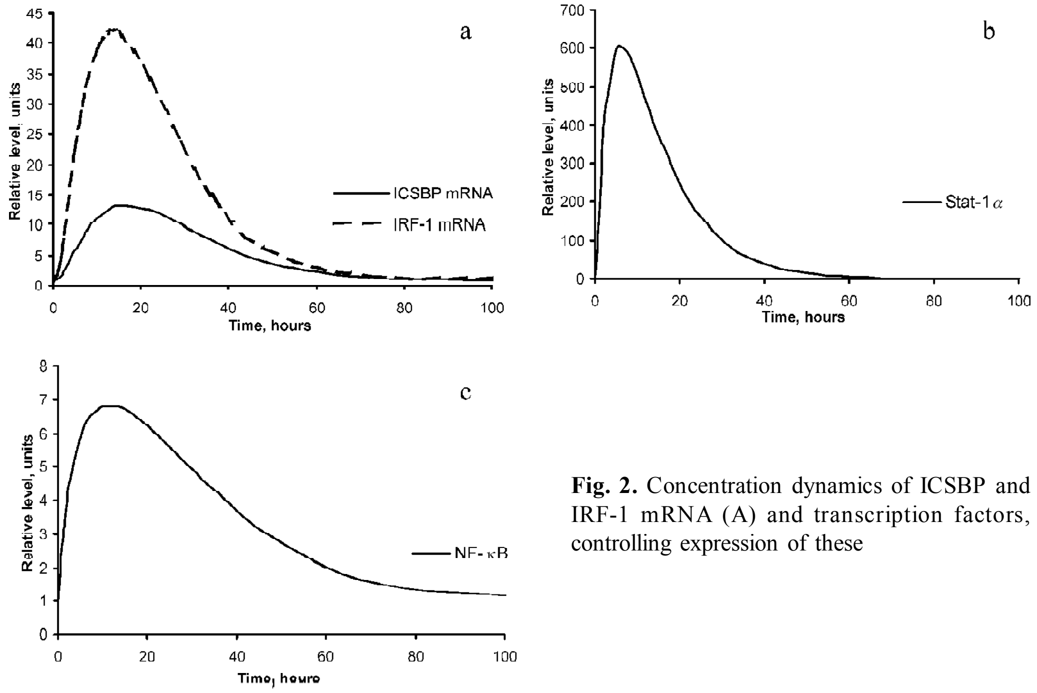


Fig. 2. Concentration dynamics of ICSBP and IRF-1 mRNA (A) and transcription factors, controlling expression of these

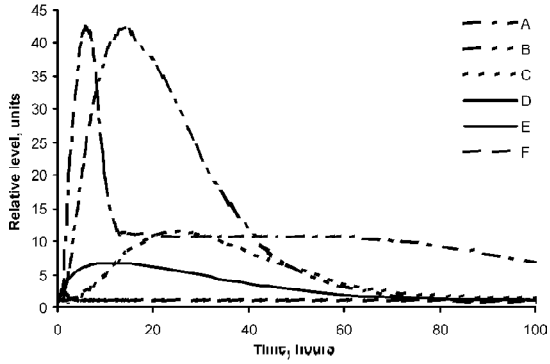


Fig. 3. Concentration dynamics of the following transcription factors upon macrophage activation by LPS: A, AP-1; B, IRF-1; C, ICSBP; D, NFκB; E, p38 MAPK; and F, Elk-1.

Discussion

The developed computer model of macrophage activation gives the presentation of the work of transcription factors upon the action of LPS on the cell. Concentrations of individual TFs reflect data obtained experimentally. In such cases, computations using the model may assist in considering macrophage activation at other LPS concentrations or in the case certain mutations are introduced into the gene network. Note that usually experiments are performed during 1–2 days; in such case, further changes in concentrations of gene network components may be found using the mathematical model. If the experimental data on quantitative characteristics of TFs are absent, the computer model can be used for calculation of these characteristics using the data on signal pathways activating TFs or the data on expression dynamics of the genes regulated by these TFs.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 02-07-90359, 02-04-48802-a, 03-04-48506-a, 03-01-00328, 03-04-48829, 03-07-96833, and 04-01-00458); NATO (grant No. LST.CLG.979816), Russian Ministry of Industry, Sciences, and Technologies (grant No. 43.073.1.1.1501), Russian Federal Research Development Program (contract No. 38/2004), and Siberian Branch of the Russian Academy of Sciences (integration projects Nos. 119, 148), and RAS Presidium Program "Molecular and Cellular Biology" (project 10.4).

References

Hambleton J., Weinstein S.L., Lem L., DeFranco A.L. Activation of c-Jun N-terminal kinase in bacterial lipopolysaccharide-stimulated macrophages // Proc. Natl Acad. Sci. USA. 1996. V. 93. P. 2774–2778.

Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan'ko E.A., Ignat'eva E.V., Podkolodnaia O.A. A generalized chemical-kinetic method for modeling gene networks // Mol. Biol. (Mosk.). 2001. V. 35. P. 1072–1079.

Likhoshvai V.A., Matushkin Yu.G., Vatolin Yu.N., Bazhan S.I. A generalized chemical kinetic method for simulating complex biological systems. A computer model of λ phage ontogenesis // Computational Technologies. 2000. V. 5. P. 87–99.

Likhoshvai V.A., Nedosekina E.A., Ratushny A.V., Podkolodny N.L. Technology of usage of experimental data for verification of the models of gene network functioning // Proc. III Intern. Conference on Bioinformatics of Genome Regulation and Structure, 2002. V. 2. P. 146–149.

Nedosekina E.A., Ananko E.A., Milanesi L., Likhoshvai V.A., Kolchanov N.A. Mathematical simulation of dynamics of macrophage gene network activated by lipopolysaccharides and/or interferon-gamma // Bioinformatics of genome regulation and structure / Ed. N. Kolchanov, R. Hofestaedt. Kluwer Academic Publishers, Boston/Dordrecht/London, 2003. P. 283–293.

Nedosekina E.A., Likhoshvai V.A., Ananko E.A. Signaling in macrophage activation gene network: data accumulation and computer modeling // Proc. of Moscow Conference on Computational Molecular Biology (MCCMB'2003). P. 159–160.

COMPUTER ANALYSIS OF THE LABELED MITOSES CURVES

Nekrasov V.*[1]*, Chernyshev A.*[1]*, Omelyanchuk L.*[2]**

[1] Institute of Chemical Kinetics SB RAS; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia

* Corresponding author: e-mail: ome@bionet.nsc.ru

Keywords: *cell cycle, mitosis, differential equation, computer analysis*

Summary

Motivation: The interpretation of the labeled mitoses curves needs improvement, since for a long time it was based on a simple model with a constant cell flow through the cell-cycle – assumption evidently incorrect now for tissues. The problem of the presence or absence of mother-daughter and daughter-daughter cell cycles correlation do not have direct experimental solution now and needs another kind of proof.

Results: The cell cycle can be described by the differential equation of wave [1]) (not to be mixed with wave equation). The comparison of the model with absolute and absent mother-daughter cell cell-cycle speed correlation shows the last one is more convenient for the cell-cycles in tissues.

Introduction

Cell-cycle is a periodic change $G1$, S, $G2$ and M phases in a cell. The method of labeled mitoses based on short time incubation of cells in labeled DNA precursors (incorporating in the S phase cells) and registration of labeled metaphase cell proportion as a function of time. [1]. If all cells have the same speed of cell-cycle, than labeled mitoses curve must to be the sequence of trapezium like figures [1]. Real curves are more smooth and the interpretation of those curves require most complex approach (a lot of other experimental facts argue against an unique cell-cycle speed [2]). Here we present a model where the cells traverse cell-cycle with a continuum of speeds distributed normally. Two extreme cases are discussed – absolute and absent mother-daughter cell-cycle correlation.

Model

The cell can be characterized at the moment t by the cell-cycle phase θ: $0 \le \theta \le 1$. $G1$, S, $G2$ phases constitute to regions $1 < \theta < \theta_{G1}$, $\theta_{G1} \le \theta < \theta_S$ and $\theta_S \le \theta < 1$, where θ_{G1}, θ_S are ending points of $G1$ and S, correspondingly. At the point $\theta_{G2} = 1$ the mitosis takes place (the length of the mitosis is usually very short in comparison to the total cell-cycle duration). As the result of the mitosis, the cell with θ=1 disappeared and two new cells with θ=0 appeared. The amount of cells, $dN(\theta,t)$, in the interval $[\theta, \theta + d\theta]$ can be expressed through the density of cells, $n(\theta,t)$: $dN(\theta,t) = n(\theta,t)d\theta$. For the density of cells the conservation of the cells flux takes place:

$$\frac{\partial n(\theta,t)}{\partial t} - \beta \frac{\partial n(\theta,t)}{\partial \theta} = 0, \tag{1}$$

where β is the cell cycle speed along the phase θ (without death). If we introduce the cell-cycle speed distribution $n(\theta,t) = n(\theta,\beta,t)$, then the equation (1) becomes:

$$\frac{\partial n(\theta,\beta,t)}{\partial t} - \beta \frac{\partial n(\theta,\beta,t)}{\partial \theta} = 0 \tag{2}$$

The general solution of equation (2) is $n(\theta,\beta,t)=f(\theta-\beta t)\rho(\beta)$, where $f(x)$ and $\rho(x)$ are some arbitrary functions. If few cell cycles take place, then the solution at the point of mitosis is

$$n(1,\beta,t)=\rho_0(\beta)\sum_{i=0}^{\infty} f_0(1-\beta t+i) . \tag{3}$$

The summation in equation (3) is over the cells that were born after *i* divisions. In order to use equation (3) one can assume the uniform phase distribution, since it is a good approximation in many cases:

$$f_0(x)=\begin{cases}0 & if \quad x\le 0\\ 1 & if \quad 0<x\le 1\\ 0 & if \quad 1<x\end{cases} . \tag{4}$$

The density of labeled cells at the point of mitosis is

$$\Phi_m(t)=\int_{\beta_{\min}}^{\beta_{\max}}\left[\rho_0(\beta)\sum_{i=0}^{\infty} f_{0m}(1-\beta t+i)\right]d\beta , \tag{5}$$

where:

$$f_{0m}(x)=\begin{cases}0 & if \quad x\le G1\\ 1 & if \quad \theta_{G1}<x\le\theta_S\\ 0 & if \quad \theta_S<x\end{cases} \tag{6}$$

the density of all the cells at the point of mitosis is

$$\Phi_n(t)=\int_{\beta_{\min}}^{\beta_{\max}}\left[\rho_0(\beta)\sum_{i=0}^{\infty} f_0(1-\beta t+i)\right]d\beta . \tag{7}$$

The ratio: $z(t)=\dfrac{\Phi_m(t)}{\Phi_n(t)}$ is the labeled mitoses curve.

In the model described above the daughter cells have the same speed β as their mother cell. It may be not the case in real experiment. It is known that a cell is sensitive to external signals at the «restriction point», θ_r, where the cell can change their speed β. The restriction point is placed within *G*1 phase near *G*1-*S* transition. In order to consider this case (speed decorrelation between mother and daughter cells) as well, we developed an alternative model assuming that at the restriction point the speed distribution for daughter cells is changed to the same initial speed distribution, $\rho_0(\beta)$, of all cells at that point. Then the distribution of the cells over cell-cycle speed at the point can be expressed as

$$\rho(\beta,\theta_r,t) = \rho_0(\beta)\sum_{i=0}^{\infty} A_i(t), \tag{8}$$

where $A_i(t)$ is the portion of cells which were born after i divisions. Due to the conservation of the cell flow for those subpopulations, the following recurrent relation takes place:

$$A_{i+1}(t) = \frac{2\int_{\beta_{\min}}^{\beta_{\max}} \beta A_i(t-\frac{1}{\beta})\rho_0(\beta)d\beta}{\int_{\beta_{\min}}^{\beta_{\max}} \beta\rho_0(\beta)d\beta}. \tag{9}$$

At the point of mitosis, the density of the cells appeared after i divisions is

$$n_i(1,\beta,t) = A_i(t-\frac{1-\theta_r}{\beta})\rho_0(\beta), \quad i>0. \tag{10}$$

In order to obtain $n_0(1,\beta,t)$, one can use equation (3) for first cell-cycle: $n_0(1,\beta,t) = f_0(1-\beta t)\ \rho_0(\beta)$. Then the following expression for $A_1(t)$ is used:

$$A_1(t) = \frac{2\int_{\beta_{\min}}^{\beta_{\max}} \beta f_0(1-\beta t)\rho_0(\beta)d\beta}{\int_{\beta_{\min}}^{\beta_{\max}} \beta\,\rho_0(\beta)d\beta}. \tag{11}$$

Thus, when the initial distribution, $\rho_0(\beta)$, (e.g. normal distribution) and phase distribution, $f_0(\theta)$, (e.g. uniform distribution) are set, one can define theoretical approximation of the labeled mitoses curve, $z(t)$. Essential assumption of both models is the independence of the relative duration of the cell cycle phases (i.e. the positions ofθ_{G1}, θ_S andθ_r) on the cell cycle speed, β.

Both models were realized as a computer programs using Labview software assuming normal initial speed distribution and uniform initial phase distribution for the cells. The optimization of the experimental and theoretical labeled mitoses curves were fitted with the variation of θ_{G1}, θ_S and the parameters of cell-cycle speed distribution. The optimization of the labeled mitoses curve for epithelial cells of duodenum of 18 days rat embryos [3] shows that the model with decorrelated mitoses describes the experimental data better than the model with absolute correlation. (χ^2 is equal to 0.0197 and 0.0233, consequently). The similar analysis was performed for a qualitatively different case: the culture of cells *in vitro*. The labeled mitoses curve for *in vitro* culturing human leucocytes where the inter cell communications is absent [4]. The optimization gives χ^2 =0.0195 and 0.0189 for the decorrelated and correlated cases, consequently. Thus our treatment demonstrated more arguments to the absence of maternal-daughter cell-cycle correlation in some tissues.

References

1. Quastler H., Sherman F.G. Cell population kinetics in the intestinal epithelium of the mouse Exp // Cell Res. 1959. V. 17(3). P. 420–438.
2. Gonzalez-Gaitan M., Capdevila M., Garcia-Bellido A. Cell proliferation patterns in the wing imaginal disc of Drosophila // Mech. Dvel. 1994. V. 40. P. 183–200.
3. Zavarzin A.A. A study of DNA synthesis and the duration of the mitotic cycle during histigenesis of the intestinal epithelium // The study of cell-cycles and nucleic acids metabolism in the cell differentiation. M.; L.: Nauka, 1964. P. 35–59.
4. Cave M. Incorporation of tritium-labeled thymidine and lysine into chromosomes of cultured human leukocytes // J. Cell Biol. 1966. V. 29(2). P. 209–222.

A COHERENT KINETIC MODEL OF SENSING AND RESPONSE IN *HALOBACTERIUM SALINARIUM* PHOTOTAXIS BASED ON THE MECHANISM OF FLAGELLAR MOTOR SWITCHING

*Nutsch T.**[1], *Marwan W.*[2], *Oesterhelt D.*[3], *Gilles E.D.*[1]

[1] Max-Planck-Institute for Dynamics of Complex Technical Systems, Magdeburg, G.; [2] Science & Technology Research Centre, University of Hertfordshire, U.K.; [3] Max-Planck-Institute for Biochemistry, Martinsried, Germany

* Corresponding author: e-mail: torsten@mpi-magdeburg.mpg.de

Keywords: *Halobacterium, taxis, motor switching, mathematical model, simulation*

Summary

Halobacterium salinarium shows a qualitatively different swimming behavior than *E. coli*, what demands a different mechanism of flagellar motor switching. In this study we postulate general properties of the switching mechanism in *Halobacteria*, derived from experimental findings and present a detailed model that quantitatively reproduces various different experimental results with the same set of parameters. Even seemingly paradox findings are accomplished by the presented model.

Introduction

Like *E. coli*, *Halobacteria* can perform tactic movements by modulating the probability of switching the rotational sense of their flagellar motors. But the swimming behavior is different in both organisms: *E. coli* performs runs and tumbles, while *Halobacterium* swims back and forth by rotating its flagella counter-clock-wise (CCW) or clock-wise (CW), respectively. The different swimming behavior demands a different switching behavior of the flagellar motors, though the switching signal CheY-P that binds to the switching complex is orthologous in both bacteria. CheY phosphorylattion is regulated by the receptor complexes in a stimulus-dependent manner. In response to a repellent stimulus, the intracellular concentration of CheY-P is increased, which in turn stabilizes the tumbly state in *E. coli*, while in *Halobacteria* the current swimming phase is terminated early (see Fig. 1). Additionally *Halobacteria* are able to sense light of different wavelengths apart from various chemo-effectors, what allows working with complex experimental stimulation programs, even at singlecell level.

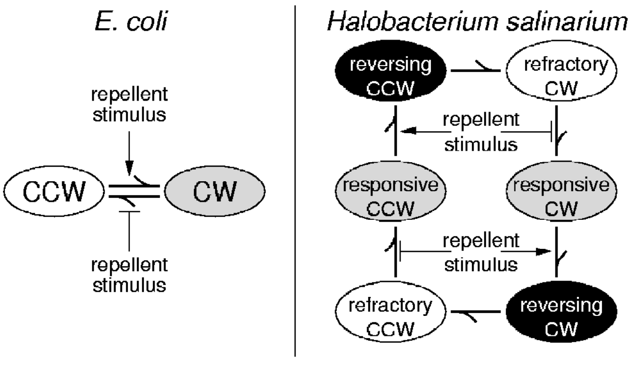


Fig. 1. General switching behavior of *E. coli* and *H. salinarium*. The model for *E. coli* is in accordance to the four-state model presented by (Kuo, Koshland, 1989) and the model for *Halobacteria* is based on (McCain *et al*., 1987).

Model

The proposed model of the switching complex is shown in Fig. 2. The exact mathematical formulation and parameters can be seen in: http://wwwa.mpi-magdeburg.mpg.de/people/torsten/switchmodel.html. The functional principle is described in the next section.

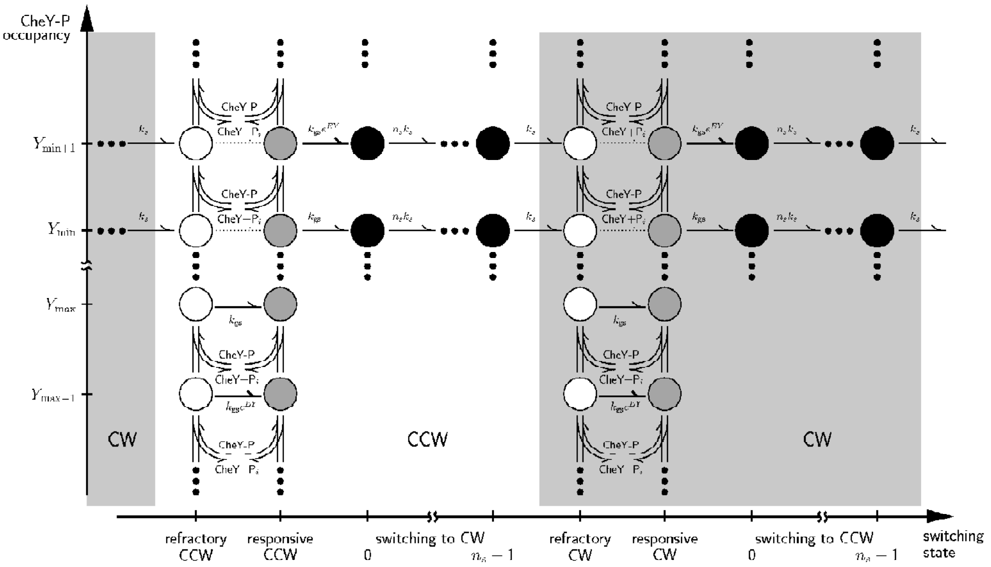


Fig. 2. Detailed model of the switching mechanism introducing a refractory, a responsive and a reversing period.

Results

We postulate the following qualitative properties of the switching mechanism of *Halobacterium salinarium* (cf. Fig. 1): a) effect of stimulus is **symmetric** in respect of rotational sense; b) mechanism must comprise a **refractory**, a **responsive** and a **reversing** period; c) switching process is an **energy** consuming, **irreversible cycle process;** d) after several **stimulus-dependent** switching steps there seems to be a number of **stimulus independent steps**.

The most apparent property of the switching mechanism is its symmetric behavior in respect to stimuli applied at either swimming direction. This symmetry property is reflected by the symmetric design of the Model (Fig. 2).

From experiments of (Krohs, 1994) it can be concluded that the switching mechanism proceeds through a **refractory period** directly after the last reversal before it enters the responsive period. Cells respond differently during refractory period than during responsive period. While given during the **responsive period** a repellent stimulus induces a quick reversal of the swimming direction with a distinct frequency distribution of the response time, the same stimulus applied during refractory period results in a delayed response with a broad response time distribution. This property is represented in our model by different transition probabilities to and from the responsive period that are assumed to depend on the occupancy level of the switching complex. In refractory period the affinity of CheY-P to the switch is assumed to be low, resulting in a net dissociation of CheY-P and in turn increasing the transition probability to the responsive period. In the responsive period, binding of CheY-P is tighter again leading to net association of CheY-P and switch complex and increasing the probability to switch to the rotational sense.

Another qualitative property can be derived from the same experimental result In *Halobacteria* the switching process must undergo an **irreversible cycle process** from refractory CW to responsive CW to refractory CCW to responsive CCW and again to refractory CW. This process must be virtually irreversible because 100 % of the cell population reacts in a delayed way to a repellent signal directly after a spontaneous reversal, what is characteristic for the refractory period, while after some seconds all cells react in the responsive prompt way, i.e. they are in the responsive period. Thus the underlying process must be irreversible, because there is no equilibration between the different states. This is a represented in the model by irreversible transitions from refractory to responsive and irreversible reversing transitions.

The third property mentioned above implies that **several rate limiting reaction steps** are necessary for the switching process. This conclusion is apparent from the frequency distribution of swimming interval length. The frequency (probability density) of a spontaneous switching event is very low for short interval lengths, but then increases steeply and finally decays exponentially again for long interval lengths (Hildebrand, Schimz, 1985) and (Marwan, Oesterhelt, 1987). If the underlying switching process would be mediated by only one reaction step, the frequency distribution would be exclusively exponential. If it would consist of two reaction steps it would increase linearly and then fall exponentially again what still doesn't reflect the experimental results. The more rate limiting reaction steps are necessary for a particular process the lower the probability for short durations of the process (first passage time) and the steeper and distinct the distribution becomes.

The underlying reaction steps of the process can be further divided into some **stimulus dependent steps** followed by a number of **stimulus independent switching steps**. Experiments of (Marwan, Oesterhelt, 1987) with single and double blue light pulses of varying intensity and duration have shown that the mean reversal time of the cells, after refractory period, is: $t_R=t_{min}+b/\tau+D\tau_2/\tau$. Where t_{min} is a constant time that is not influenced by the stimulus. The reversal time t_R is proportional to the total light pulse $I_{bl}\tau$ (b is a constant) and is delayed by the dark period D weighted by the ratio of the second pulse τ_2 and the total pulse τ. The main statement of this formula is that the second blue light pulse still has the same effectiveness (proportional to its duration) in inducing the reversal event as the first one, only that its impact is delayed by the dark period D. Thus the minimal time t_{min} can not result from nonlinear saturation effects, but a separate stimulus independent process has to be considered (Nutsch *et al.*, 2003). This process must consist of numerous necessary

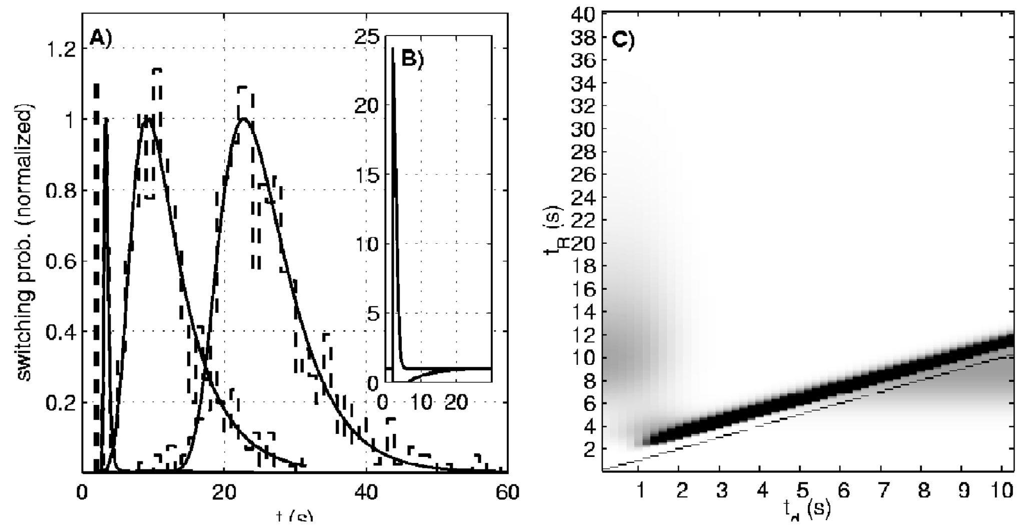


Fig. 3. (A) frequency distribution of interval lengths in the repellent induced (maximum at 3s), spontaneous (maximum at 10s), and attractant induced (maximum at 25s) case. Solid lines: simulations; dashed lines: measurements taken from (Hildebrand, Schimz, 1985). Dashed line at t = 2s indicates start of repellent and attractant stimulus; (B) normalized CheY-P concentration for repellent (positive peak) and attractant (negative peak); (C) Simulation results of response to repellent stimuli (UV step-up) in dependence of the delay of stimulation during an interval. For each delay t_d the normalized frequency distribution of reversals is represented in grayscales along the y-axis. The bisecting line shows the time point of stimulation.

reaction steps, because the frequency distribution of the response time measured after applying a strong repellent light stimulus shows a very narrow peak (Hildebrand, Schimz, 1985).

The **Simulation Results** are shown in Fig. 3. Not only the mean reversal time, but also the frequency distribution of interval lengths is described by the model in good agreement with the experimental findings for spontaneous, repellent and attractant stimuli. Also the transition from refractory period to responsive period (Fig. 3C) is well made by our model while keeping the set of parameters. For simulating the response to attractant stimuli we had to introduce an adaptational mechanism into the kinetic model of the excitation pathway, because the reversal time upon to attractant stimulation is in the same range as the adaptation time. Details of the adaptation mechanism are not given here, although the time-course of CheY-P concentration in the repellent and attractant stimulated case is shown in Fig. 3B.

In addition, seemingly paradox experimental results that were published in (McCain *et al.*, 1987) are described correctly reproduced by our model (simulations not shown): An orange light pulse of 1s normally induces an attractant response, i.e. the suppression of reversals. But the same pulse induces reversals when it is applied 2 to 6s after a repellent stimulus. The following mechanism is proposed by our model: The CheY-P concentration is increased by the repellent stimulus at the beginning of the experiment, what prolongs the refractory period to approximately 8s. A short attractant light pulse during this time transiently reduces the CheY-P concentration, what induces the transition to responsive period. After the attractant light pulse the CheY-P concentration rises again and initiates the next reversal.

Discussion

The switching behavior of *Halobacterium salinarium* as measured in various types of stimulation experiments is described for the first time by a new model for the switching mechanism. The same set of parameters could be used throughout and no complex and mechanistically questionable time-courses of CheY-P concentration (Naber, 1997) had to be assumed. In addition, the experimental results that have led to the assumption of an oscillator-driven switching mechanism (Schimz, Hildebrand, 1985) can be reproduced by our model (simulations not shown).

The kinetic scheme in Fig. 1 can be regarded as a parent model (McCain *et al.*, 1987), which qualitatively explains the above-mentioned experimental results while the of the detailed model, containing more kinetic states (Fig. 2) quantitatively reproduces all experimental observations analyzed in this respect until today. Hence we propose the first coherent kinetic model of sensing and response in prokaryotic signaling.

References

Hildebrand E., Schimz A. Sensing and Response in Microorganisms. Elsevier Science Publischer B.V., 1985. P. 129–142.

Krohs U. Sensitivity of *Halobacterium* s. to attractant light stimuli does not change periodically // FEBS. 1997. V. 351. P. 133–136.

Kuo S.C., Koshland Jr. D.E. Multiple kinetic states for the flagellar motor switch // J. Bacteriol. 1989. V. 171. P. 6279–87.

Marwan W., Oesterhelt D. Signal Formation in the *Halobacterial* Photophobic Response Mediated by a Fourth Retinal Protein (P475) // J. Mol. Biol. 1987. V. 195. P. 333–342.

McCain D.A., Amici L.A., Spudich J.L. Kinetically resolved states of the *Halobacterium* halobium flagellar motor switch and modulation of the switch by sensory rhodopsin I // J. Bacteriol. 1987. V. 169. P. 4750–4758.

Naber H. The response of halobacteria to single light stimuli: a theoretical analysis // Eur. Biophys. J. 1997. V. 26. P. 163–173.

Nutsch T., Marwan W., Oesterhelt D., Gilles E.D. Signal processing and flagellar motor switching during phototaxis of *Halobacterium salinarum* // Genome Res. 2003. V. 13. P. 2406–2412.

AN ELEMENTARY MODULE RECOGNIZING MORPHOGENETIC GRADIENTS IN TISSUES

Omelyanchuk L.V.*, Gunbin K.V.

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

Keywords: *pattern formation; positional information; molecular trigger; mathematical model*

Summary

Motivation: Analysis of complex gene networks dictates subdivision into simper tasks. In the present study, our aim was to reveal the simplest molecular machinery responsible for interpretation of the morphogene gradients.

Results: The simplest chemical-kinetics mathematical model of elementary machinery responding to the gradient of the HH morphogene at the anterior/posterior (A/P) compartment boundary of the *Drosophila melanogaster* wing imaginal disc was constructed. The model simulates the expected threshold behavior of the PTC protein expression at the A/P compartment boundary.

Introduction

Wolpert has suggested the pattern formation concept also known as the “positional information” hypothesis (Wolpert, 1969). The hypothesis considers the development of morphological structures as a sequence of events resulting in a morphogene gradient, wich is interpreted by the cells activating other different genes depending on morphogene concentration. Current knowledge about development shows the concept is reasonable (Held, 2002). Computer modeling of embryonic segmentation in *D. melanogaster* demonstrates that the current data on the gene network may be sufficient or almost so for understanding why periodic gene expression patterns appear in the early embryo; to describe how the morphogenetic network functions, chemical kinetics equations (distributed along a coordinate) with stochastically generated kinetics constants have been applied, and the real pattern of the segmentation genes expression in the embryo has been simulated (Dassow *et al.*, 2000). Despite its advantages the model (Dassow *et al.*, 2000) does not provide opportunities for separating the behavior of morphogene gradients from the cell molecular machinery recognizing the gradients.

Here, our analysis of the gene network of the *D. melanogaster* A/P boundary formation identified, the machinery for alternative regulation of the *ptc* gene expression whose product is the morphogene HH receptor, the PTC protein binds to HH, thereby inducing the degradation of HH. It is of importance that when in a state not bound to HH, the PTC protein induces the formation of the inhibitory form of the CI protein, which suppresses *ptc* transcription. When the PTC-HH complex is generated, the PTC activity is suppressed so that the formation of the activator CI starts. This activates *ptc* transcription. Thus, the behavior of the system can be described as a molecular trigger recognizing the morphogene gradient. Here, this system is qualitatively analyzed using chemical-kinetics equations.

Model

A scheme for the kinetics of the elementary machinery presumably containing the “molecular trigger” is outlined in Fig. 1. The *ptc* gene expression regulation is the most important unit of the block responding to the HH concentration gradient. The transcriptional activity of the *ptc* gene is dependent on the ratio of the concentrations of the activator form of the CI transcription factor (Ci(a)) to the inhibitor form Ci(i). Therefore, the transcriptional activity of the *ptc* gene was expressed as $k_0Ci(a)/(k_1+k_2Ci(i))$. It was assumed that protein Ci(i) competed with the Ci(a) protein during *ptc* transcription for the enhancer site. The Ci(a) protein, the *ptc* gene and the other units of the basal transcription complex are considered as an enzyme. For this reason, the reaction rate is a linear, not a hyperbolic function of the enzyme Ci(a) concentration (Cornish-Bowden, 1979). It was assumed that the translation of the *ptc* gene was a monomolecular reaction with the rate constant k_3. The PTC protein can degrade with the constant kd_1 and/or associate with the morphogene HH with the rate constant k_4 to form the PTC-HH complex at the cell surface. When not bound to HH, the PTC protein inhibits the activity of the SMO protein and promotes the formation of the Ci(i) form: $Ci(i)$=β[PTC]. When in the free state, not inhibited through PTC, the SMO protein facilitates the formation of the activator form Ci(a). It was assumed that the SMO concentration remained constant and, consequently, the Ci(a) concentration depended on the PTC concentration: $Ci(a)$=α[PTC]. As a result, the set of differential kinetics equations for the block responding to the HH concentration gradient became:

$$\frac{dx_1}{dt} = \frac{k_0 \cdot Ci(a)}{k_1 + k_2 \cdot Ci(i)} - k_3 \cdot x_1$$

$$\frac{dx_2}{dt} = k_3 \cdot x_1 - k_4 \cdot x_2 \cdot HH - kd_1 \cdot x_2$$

$$Ci(a) = \alpha \cdot x_2$$

$$Ci(i) = \beta \cdot x_2$$

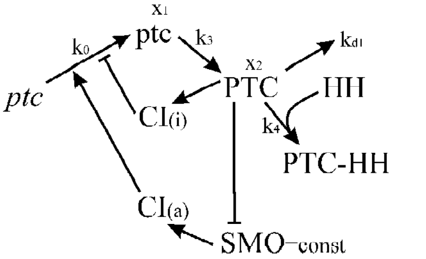


Fig. 1. A schematic representation of the kinetics for the block responding to the HH concentration gradient at the anterior/posterior (A/P) compartment boundary in the *Drosophila melanogaster* wing imaginal disc.

Here, x_1 denotes the concentration of the *ptc* mRNA, x_2 is the concentration of the PTC protein at the cell surface. Compared with transcription, morphogene diffusion or the formation of the protein complexes, morphogenesis is much slower, making permissible analysis of the stationary state:

$$\frac{k_0 \cdot Ci(a)}{k_1 + k_2 \cdot Ci(i)} = k_3 \cdot x_1$$

$$k_4 \cdot x_2 \cdot HH + kd_1 \cdot x_2 = k_3 \cdot x_1$$

and

$$Ci(a) = \alpha \cdot x_2, \quad Ci(i) = \beta \cdot x_2.$$

In this case, the concentration of the PTC protein (x_2) is easily found proceeding on the conditions of the following set of equations:

$$x_2 = A \cdot \frac{(HH - B)}{(C + HH)}$$

$$A = -\frac{1}{k_2 \beta}$$

$$B = \frac{\alpha \cdot k_0 - k_1 \cdot kd_1}{k_1 \cdot k_4}$$

$$C = \frac{kd_1}{k_4}.$$

In the *D. melanogaster* wing imaginal discs, cells of the posterior compartment form the HH morphogene, which migrates to the anterior compartment, giving rise to the HH gradient along the A/P compartment boundary. The *ptc* expression pattern appears stripe-like along the A/P boundary, with maximum expression at the sites where HH concentrations are highest (Fig. 2a).

This feature of the *ptc* expression pattern is predictable with our scheme. The qualitative behavior of the PTC concentration as a function of the HH concentration is shown in Figure 2b. Obviously, the solution of the set of equations with positive PTC concentration values is feasible only in the case if the HH concentration is greater than the *B* value. If the HH concentration is smaller than *B*, the solutions of the equation set yield zero concentrations for the variables of the set. In the range of higher HH concentrations, the solution becomes comparable to that of the real PTC protein concentration gradient (at the left to the HH=50 point, Fig. 2b). However, the set of equations is inapplicable to a description of the distribution of the PTC protein concentration in the posterior wing compartment (at the right to the HH=50 point, Fig. 2b). This is because the product of the selector gene *en,* which is expressed at the posterior compartment, inhibits the expression of the *ptc* and *ci* genes.

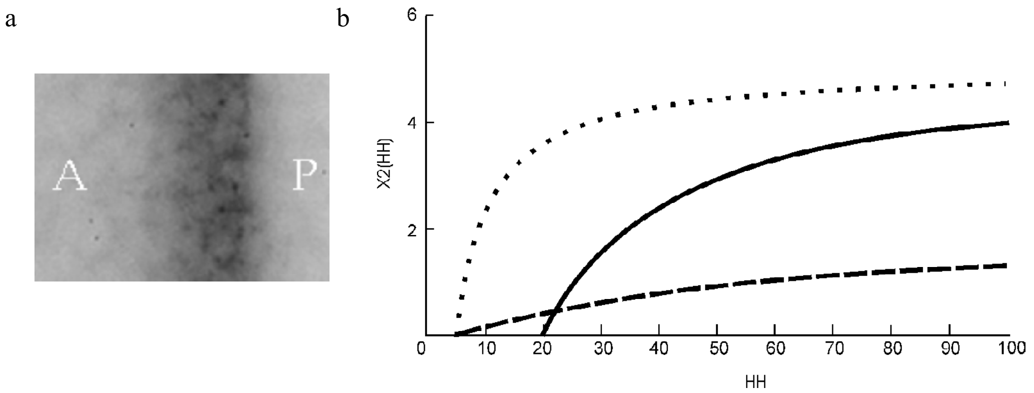


Fig. 2. a – the expression of the *ptc* gene. The letter "A" designates the anterior compartment, "P" the posterior compartment. The darker strip represents the site where the *ptc* gene expresses. The PTC gradient concentration is seen in the stripe, its minimum is in the anterior compartment (see text); b – the behavior of the function describing the PTC concentration (x2) whose argument is the HH concentration at the A/P boundary of the *D. melanogaster* imaginal disc compartments. It is assumed that the anterior compartment is at the left to the HH=50 point, and the posterior compartment is at the right to this point (see text). The concentration of the PTC protein depends on the *A*, *B* and *C* constants (the calculations were done within the framework of the model): the continuous line corresponds to *A*=5, *B*=20, *C*=1; the dotted line to – *A*=5, *B*=5, *C*=1; the dashed line to – *A*=2, *B*=5, *C*=50.

Acknowledgements

The work was supported by the RAS Presidium Program "Molecular and Cellular Biology" (Projects No.10.4) and Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501).

References

Cornish-Bowden A. Principles of Enzyme kinetics. M.: Mir, 1979.

Dassow G., Meir E., Munro E., Odell G. The segment polarity network is a robust developmental module // Nature. 2000. V. 406. P. 188–192.

Held L.I. Imaginal discs: the genetic and cellular logic of pattern formation. Cambridge: Cambridge University Press, 2002.

Wolpert L. Positional information and the spartial pattern of cellular differentiation // J. Theor. Biol. 1969. V. 25. P. 1–47.

AGNS: ARABIDOPSIS GENENET SUPPLEMENTARY DATABASE

Omelianchuk N.A.*, Mironova V.V., Poplavsky A.S., Kukeeva Yu.A., Podkolodny N.L., Kolchanov N.A.

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

Keywords: *Arabidopsis, gene expression, phenotype, database*

Summary

Motivation: In the past few years the number of publications related to phenotypic abnormalities and expression of the Arabidopsis genes in mutant and transgenic plants has greatly increased. This disparate and scattered information requires integration and systematization by modern information technologies.

Result: To resolve the problem, we developed the relational AGNS (the Arabidopsis GeneNet supplementary database), which consists now of three databases, the Expression Database (ED), the Phenotype Database (PD), the Reference Database (RD), and two controlled vocabularies. The ED describes gene expression in wild type, mutant and transgenic plants. The PD contains information on phenotypic abnormalities in mutant and transgenic plants. The RD contains references to the papers together with a description of plant growth conditions with an indication of the ecotypes used as control in the experiments. AGNS makes possible search for genes expressed in particular organs, at particular stages, for genes whose expression is altered in particular mutants, and for alleles causing similar phenotypic abnormalities.

Availability: http://vampire.bionet.nsc.ru:8080/agnsdb/.

Introduction

To study development in Arabidopsis, access is required to the various and scattered information in publications found as descriptions of gene expression patterns and morphological abnormalities in different mutant and transgenic lines compared with those for wild type. Integration of this growing body of data and its description in a common format would considerably facilitate analysis of the information. A number of databases are presently available via the Internet where information on Arabidopsis is stored. In addition to the nucleotide and protein sequence databases the EMBL (Kulikova *et al.*, 2004), the GenBank (Benson *et al.*, 2003), the SWISS-PROT (Apweiler *et al.*, 2004) and others, data on the Arabidopsis genome are also amassing in specialized databases, such as TAIR (Rhee *et al.*, 2003), TIGR (Ouyang *et al.*, 2004) and MATDB (Schoof *et al.*, 2004). Regrettably, all the databases contain very scant data related to the detailed description of the experimental results and observed expression of the Arabidopsis genes at the levels of transcription, translation, protein interactions and phenotype. The aim of AGNS is to create an Internet available resource accumulating these data from annotations of published papers and thereby, to provide a description of the functions of the Arabidopsis genes at different levels of mRNA, protein, cell, tissue and ultimately at the levels of the organ and organism and in different genotypes.

Methods and Algorithms

Web resource can be viewed with Netscape Navigator 6.0 (IE Explorer 5.0, Mozilla 1.0) or higher.

Results and Discussion

Controlled vocabulary. Regrettably, the current description of the stages, organs and tissues in Arabidopsis are found in separate papers. Most papers concern only certain stages and/or only some their aspects. This made difficult annotation of papers and data systematization. To avoid this challenge, we created controlled vocabularies, which contain information on:

1. Description of organs, tissues and cells both in the mature plant and at different developmental stages.

2. Developmental stages of the plant itself and of its separate organs and also changes in the anatomy and morphology of these organs characterizing the transition from one stage to another.

All these data have references to the papers from which they were annotated. Thus, controlled vocabularies contain information on the available descriptive systems of Arabidopsis morphology and development, which is systematized and compared. The most frequently used names of the stages, organs are highlighted and their synonyms are given. Every description of stages and organs is accompanied by detailed commentaries. All the vocabularies are supplemented with new research data as they become available.

Description of phenotypic abnormalities. The experimental papers contain rather scattered information describing changes in the phenotypes of mutant and transgenic plants. For a formalized description of data of this kind, we developed the PD, which provides collection, primary processing and classification of the data on the phenotypic effect of particular mutations during the development of particular organs. The PD contains information on abnormalities of particular organs at certain stages under the effect of a mutation or a transgenic construct (Fig.). An organ is described in a hierarchical form for search convenience. There are extensive commentaries specifying the abnormality, which can be also further used for a more detailed classification of a particular abnormality as data keep increasing, for highlighting new fields of description or subdividing the abnormalities into groups. The PD provides the user with the opportunity to obtain data on double, triple mutations, and this can help him to make conclusions about gene interaction. The PD makes possible the following automated queries, which help the user to find the needed information in the AGNS:

1. Mutations resulting in phenotypic abnormalities of the selected organs.

2. Mutant phenotypes.

Description of gene expression patterns. Gene expression may be markedly different, depending on organ, tissue, stage and action of external factors. The expression is also modified under the effect of mutations in the other genes (for example, transcription factors). Information on the expression of genes involved in the development of Arabidopsis is accumulating in the ED. This database is subdivided into two parts: one for describing the expression in the wild type and the other for its change under the effect of mutation or in a transgenic plant (an increase or a decrease in gene expression level, change in domain size or timing of expression). An important feature of data display in the ED is an obligatory reference to the stage and organs (tissues, cells) where gene expression was identified. In description of the expression blocks, there are special fields where the level of expression (RNA, protein) and the identification method (in situ or blot hybridization, RT-PCR, chip or other methods) are indicated. Thus, the ED contains data on gene expression at the levels of transcription, translation, protein interaction and on its changes in mutant and transgenic plants. Based on the ED the following automated queries are provided:

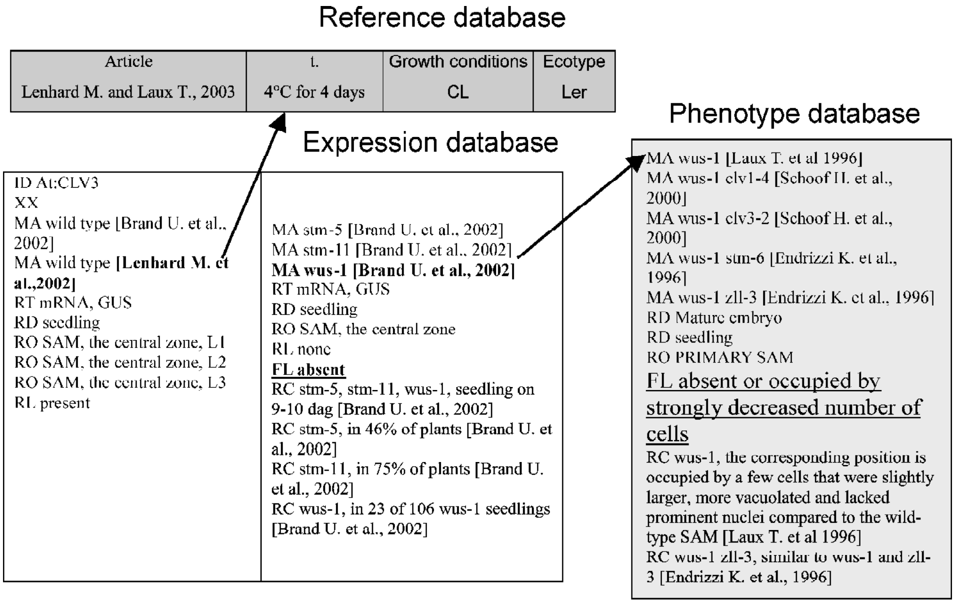
1. Expression pattern of the gene.

2. Genes expressed in certain organs.

3. Genes expressed at the queried stage.

4. Genes with abnormal expression in mutant or transgenic plants.

Reference database. It contains references to papers with a link in PubMed to which the data (Fig.) for growth conditions of plants in experiments and the name of the ecotype used as wild type are added. This data are required for studying temperature or light sensitive mutations and also mutations in different ecotypes.



Reference database

Article	t.	Growth conditions	Ecotype
Lenhard M. and Laux T., 2003	4°C for 4 days	CL	Ler

Expression database

ID At:CLV3
XX
MA wild type [Brand U. et al., 2002]
MA wild type **[Lenhard M. et al.,2002]**
RT mRNA, GUS
RD seedling
RO SAM, the central zone, L1
RO SAM, the central zone, L2
RO SAM, the central zone, L3
RL present

MA stm-5 [Brand U. et al., 2002]
MA stm-11 [Brand U. et al., 2002]
MA wus-1 [Brand U. et al., 2002]
RT mRNA, GUS
RD seedling
RO SAM, the central zone
RL none
FL absent
RC stm-5, stm-11, wus-1, seedling on 9-10 dag [Brand U. et al., 2002]
RC stm-5, in 46% of plants [Brand U. et al., 2002]
RC stm-11, in 75% of plants [Brand U. et al., 2002]
RC wus-1, in 23 of 106 wus-1 seedlings [Brand U. et al., 2002]

Phenotype database

MA wus-1 [Laux T. et al 1996]
MA wus-1 clv1-4 [Schoof H. et al., 2000]
MA wus-1 clv3-2 [Schoof H. et al., 2000]
MA wus-1 stm-6 [Endrizzi K. et al., 1996]
MA wus-1 zll-3 [Endrizzi K. et al., 1996]
RD Mature embryo
RD seedling
RO PRIMARY SAM
FL absent or occupied by strongly decreased number of cells
RC wus-1, the corresponding position is occupied by a few cells that were slightly larger, more vacuolated and lacked prominent nuclei compared to the wild-type SAM [Laux T. et al 1996]
RC wus-1 zll-3, similar to wus-1 and zll-3 [Endrizzi K. et al., 1996]

Fig. Fragment of the database composing the AGNS. A description of CLV3 expression in the Expression Database in wild type and mutants (the designations are MA, alleles; RT, detection method; RD, stage; RO, organ in which expression was identified; RL, level of expression; FL, abnormality; RC, commentaries) and SAM abnormality caused by one of the mutations (Phenotype Database). Also shown is the link on the Reference database with an indication of the growth conditions of plants used in the experiments. The relation among the databases is shown.

Information content of the AGNS. At present, the expression of 109 genes is described on the basis of 182 papers in the AGNS. The phenotypic abnormalities of mutants and also the expression of the genes are most completely examined in the shoot apical meristem.

Conclusions. Thus, a feature of the AGNS database is detailed annotations of published experiments and observations on the expression of the Arabidopsis genes, and also on the regulation of the expression at the levels of transcription, translation, protein interactions and the phenotype. These data may be useful to a wide range of researchers in the area of plant genetics and development, furthermore controlled vocabularies may be used both for explanation and comparison of data from the AGNS and may be a curated gateway to the various descriptions of the stages and synonyms of the Arabidopsis organs. Automated queries provide an access to the component of the database directly relevant to user interest. Further development of the AGNS database will be aimed at improving its format. This will also allow taking into consideration the data on changes in nucleotide sequences in mutant alleles and also on the structure and function of the Arabidopsis proteins.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (grant Nos. 02-07-90359, 03-07-96833, 03-04-48506-a, 03-07-90181-в, 03-04-48469-a), RAS Presidium Program “Molecular and Cellular Biology” (grant No. 10.4), Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501), Siberian Branch of Russian Academy of Sciences (Integration Projects No. 119), and US National Science Foundation FIBR PR 03-106 “Developing modeling and bioinformatics”.

Reference

Apweiler R., Bairoch A., Wu C.H. Protein sequence databases // Current Opinion in Chemical Biol. 2004. V. 8. P. 76–80.

Benson D.A., Karsch-Mizrachi I., Lipman D.J., Ostell J., Wheeler D.L. GenBank // Nucleic Acids Res. 2003. V. 31, N 1. P. 23–27.

Kulikova T., Aldebert P., Althorpe N., Baker W., Bates K., Browne P., van den Broek A., Cochrane G., Duggan K., Eberhardt R., Faruque N., Garcia-Pastor M., Harte N., Kanz C., Leinonen R., Lin Q., Lombard V., Lopez R., Mancuso R., McHale M., Nardone F., Silventoinen V., Stoehr P., Stoesser G., Tuli M.A., Tzouvara K., Vaughan R., Wu D., Zhu W., Apweiler R. The EMBL nucleotide sequence database // Nucleic Acids Res. 2004. V. 32. Database issue:D27–30.

Ouyang S., Buell C.R. The TIGR plant repeat databases: a collective resource for the identification of repetitive sequences in plants // Nucleic Acids Res. 2004. V. 32. Database issue D360–D363.

Rhee S., Beavis W., Berardini T.Z., Chen G., Dixon D., Doyle A., Garcia-Hernandez M., Huala E., Lander G., Montoya M., Miller N., Mueller L.A., Mundodi S., Reiser L., Tacklind J., Weems D.C, Wu Y., Xu I., Yoo D., Yoon J., Zhang P. The Arabidopsis information resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community // Nucleic Acids Res. 2003. V. 31, N 1. P. 224–228.

Schoof H., Ernst R., Nazarov V., Pfeifer L., Mewes H., Mayer K.F.X. MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics // Nucleic Acids Res. 2004. V. 32. (Database issue): D373–D376.

PROTEIN-PROTEIN INTERACTION NETWORK OF APOPTOSIS SIGNALING PATHWAYS AND ITS USAGE TO IDENTIFY APOPTOTIC REGULATORY ELEMENTS

Peng F.*, Pio F.

Department of Molecular Biology & Biochemistry, Simon Fraser University, Burnaby, British Columbia, Canada
* Corresponding author: e-mail: ypeng@sfu.ca

Keywords: *protein-protein interaction, transcription factor binding site (TFBS), regulatory networks, phylogenetic footprinting, TFBS Perl system, LAGAN, VISTA, apoptosis*

Summary

Motivation: Aberrant regulation of apoptosis is associated with multiple human diseases. Understanding gene regulatory networks and protein-protein interactions of apoptotic signaling cascades may help us identify novel biomarkers and drug targets for apoptosis-related diseases such as cancers.

Results: We have assembled a protein-protein interaction network of genes involved in apoptotic signaling pathways, and used them together with phylogenetic footprinting between human and mouse to identify putative transcription factor binding sites controlling apoptosis. Our results reveal that this approach can retain fairly good prediction accuracy while reducing the false positive rates significantly.

Availability: All scripts and the entire network are freely available upon request.

Introduction

Apoptosis is a form of cell death that removes damaged or unnecessary cells. It plays critical roles in controlling cell population during development (Ashe, Berry, 2003; Lawen, 2003). However, aberrant regulation of apoptosis is the pathological cause of many human disorders, such as neurodegenerative disorders and cancers. Thus, intensive studies on regulation of apoptosis signaling cascades have been performing to understand the mechanisms of various diseases and to identify novel drug targets. Reed J.C. *et al.* (2003) identified over 200 apoptosis genes in each of human and mouse genomes. We are analyzing the promoter regions of these genes to understand their transcriptional regulation in apoptotic signaling pathways by using the TFBS Perl system (Lenhard, Wasserman, 2002). The predictions are filtered by phylogenetic footprinting and protein-protein interaction data.

Methods

1. The apoptosis genes in human and mouse were compiled by Reed *et al.* (2003), their accessions were retrieved from NCBI. The transcript accessions were used to obtain their 1 kb upstream sequences from UCSC. Known transcription factors and binding matrices were from the TRANSFAC database (Wingender *et al.*, 2001). Protein interactions were extracted from human protein reference database (Peri *et al.*, 2003), and the network was built using Cytoscape (Shannon *et al.*, 2003).

2. TFBSs are identified in a 3-step approach. First, the TFBS software system (Lenhard, Wasserman, 2002) was used to scan promoter regions using 507 known binding matrices from vertebrates. Second, retain sites in conserved regions between human and mouse, determined by LAGAN alignment (Brudno *et al.*, 2003) and VISTA (Mayor *et al.*, 2000). The third step is to keep sites shared in interacting genes using our protein interaction network. The prediction assessment uses

a test dataset comprised of 16 human or mouse apoptotic genes with known 37 TFBSs from TRANSFAC database, and sensitivity and specificity calculations follow Lenhard B. *et al*. (2003).

Results

We extracted 364 distinct interactions covering 160 apoptosis genes and built an interaction network. A sub-network that contains Caspase-8 interactions is shown in Fig. The entire network itself is valuable and could help biologists to come up with testable hypothesis of novel interactions or complexes in the apoptotic signaling cascades.

Table. Prediction test using protein-protein interaction and phylogenetic footprinting

Matrix threshold	70 %		80 %		85 %		90 %		95 %	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
1	81.8	628	81.8	107	81.8	41	66.7	14	42.4	4
2	75.8	165	75.8	28	72.7	10	51.5	4	36.4	1
3	75.8	490	75.8	69	75.8	26	60.6	8	30.3	2
4	81.8	465	81.8	79	81.8	29	66.7	10	42.4	2
5	69.7	169	69.7	18	66.7	7	45.5	2	24.2	0.5

Sensitivity (Sn) is percent correct predictions in the test set; Specificity (Sp) is the number of predictions in a 100 bp promoter sequence in both strands. 1=prediction without filtering; 2=only sites in the conserved regions; 3=only sites shared by interacting genes; 4=(2) OR (3); 5=(2) AND (3).

We went one step further and used the interaction network for TFBS identification (Table). The 85 % matrix threshold appears to be the best setting, while the sensitivity is the same as that of 70 % matrix threshold, the specificity is more reasonable. Less stringent threshold can only dramatically increase false positives. Strikingly, the results show that the sensitivity of using interaction data is even better than that phylogenetic footprinting under all matrix thresholds, suggesting that some TFBSs are not in the conserved regions. Furthermore, by selecting sites either in the conserved regions or shared by interacting genes, the sensitivity is fully recovered but the specificity drops compared to the prediction without filtering, indicating that many false positives are eliminated.

Discussion

A "renowned" issue in computational TFBS identification is the extremely high false positive rate, and thus other information such as sequence conservation and gene expression data must be used as additional signals. Microarray data have already been used for identifying regulatory elements. Here we show that protein-protein interaction data can also be used to improve TFBS predictions. Our assumption is that interacting genes are likely share similar transcription factors. Indeed, it was shown that the *cis*-similarity, defined as the proportion of shared TFBSs is higher for interacting proteins as well as for members of a signal transduction pathway (Hannenhalli, Levy, 2003). Our results suggest that protein-protein interaction data may complement the phylogenetic footprinting information in TFBS detection. With more high-throughput proteomics from ongoing proteomics efforts, we believe that protein-protein interaction data, like microarray data, will help us further improve TFBS predictions and eventually understand gene regulatory networks underlying many critical cellular pathways.

Acknowledgements

This work has been funded by the NSERC Operating and Genomics Grant, F. Peng is grateful to the CIHR, MSFHR and Alfred P. Sloan Foundation Bioinformatics Graduate Program for supporting this work. F. Pio is an ASI Provincial Research Fellow.

References

Ashe P.C., Berry M.D. Apoptotic signaling cascades // Progress in Neuro-Psychopharmacology & Biological Psychiatry. 2003. V. 27. P. 199–214.

Brudno M., Do C.B., Cooper G.M., Kim M.F., Davydov E., Green E.D., Sidow A., Batzoglou S. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA // Genome Res. 2003. V. 13. P. 721–731.

Hannenhalli S., Levy S. Transcriptional regulation of protein complexes and biological pathways // Mammalian Genome. 2003.V. 14. P. 611–619.

Lawen A. Apoptosis – an introduction // BioEssays. 2003. V. 25. P. 888–896.

Lenhard B., Wasserman W.W. TFBS: Computational framework for transcription factor binding site analysis // Bioinformatics. 2002. V. 18. P. 1135–1136.

Lenhard B., Sandelin A., Mendoza L., Engstrom P., Jareborg N., Wasserman W.W. Identification of conserved regulatory elements by comparative genome analysis // J. Biol. 2003. V. 2. P. 13.1–13.11.

Mayor C., Brudno M., Schwartz J.R., Poliakov A., Rubin E.M., Frazer K.A., Pachter L.S., Dubchak I. VISTA: visualizing global DNA sequence alignments of arbitrary length // Bioinformatics. 2000. V. 16. P. 1046–1047.

Peri S., Navarro J.D., Amanchy R., Kristiansen T.Z., Jonnalagadda C.K., Surendranath V., Niranjan V., Muthusamy B., Gandhi T.K.B., Gronborg M. *et al.* Development of human protein reference database as an initial platform for approaching systems biology in humans // Genome Res. 2003. V. 13. P. 2363–2371.

Reed J.C., Doctor K., Rojas A., Zapata J.M., Stehlik C., Fiorentino L., Damiano J., Roth W., Matsuzawa S., Newman R. *et al.* Comparative analysis of apoptosis and inflammation genes of mice and humans // Genome Res. 2003. V. 13. P. 1376–1388.

Shannon P., Markiel A., Ozier O., Baliga N.S., Wang J.T., Ramage D., Amin N., Schwikowski B., Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks // Genome Res. 2003. V. 13. P. 2498–2504.

Wingender E., Chen X., Fricke E., Geffers R., Hehl R., Liebich I., Krull M., Matys V., Michael H., Ohnhauser R. *et al.* The TRANSFAC system on gene expression regulation // Nucleic Acids Res. 2001. V. 29. P. 281–283.

ON RESEARCH INTO HYPOTHETICAL NETWORKS OF ECOLOGICAL NATURE

***Peshkov I.M.*[4], *Likhoshvai V.A.*[1,3], *Matushkin Yu.G.*[1], *Fadeev S.I.**[2]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Institute of Mathematics SB RAS, Novosibirsk, Russia; [3] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia; [4] Novosibirsk State University, Novosibirsk, Russia
* Corresponding author: e-mail: fadeev@math.nsc.ru

Keywords: *hypothetical networks of ecological nature, mathematical model, computer model, regulation, negative feedback, positive feedback, critical point, limit cycles, stability*

Summary

Motivation: Analysis of the influence of the structure of ecological networks on their functional properties, in particular, stationary modes of existence, is an important problem in bioinformatics. Study of the properties of theoretical constructs is a necessary stage in understanding of the function patterns of natural ecological networks.

Results: This work reports the results obtained investigating a mathematical model of hypothetical symmetrical ecological networks – the model $E_1(n,k)$. It was proved that additional stationary solutions (compared with $M_1(n,k)$, the model of hypothetical gene networks) are Lyapunov unstable at $t > 0$. Qualitative and numerical studies of the model $E_1(n,k)$ confirmed the statements of the known (n,k)-criterion formulated for the model $M_1(n,k)$. The necessary and sufficient conditions were found for the stationary solution of equation with a retarded argument (the model $E_1(n,2)$, with odd n) to be asymptotically stable. The ranges for parameter variations wherein this solution comes to a periodical mode were determined.

Introduction

Characteristic of ecological systems is an intricate scheme of interactions between their elements. One of the key interaction factors is competition and, as a consequence, a mutual suppression, either direct or indirect. The suppression appear as a competition for food or space; extermination of a species by another by predation (a trophic pyramid), or poisoning of a species by metabolites of another species. We considered a similar situation under condition of a temporal mismatch (shift) of the developmental stages of closely related (i.e., virtually similar in the parameters considered) species, when each of the species suppress the earlier developmental stages of several other species by its metabolites. Then, under conditions of a transient time, we have the situation of successive cyclic mutual inhibition of reproduction for a population of several asynchronous species.

Let us have *n* species and $n \geq k > 1$. Let us assume that the *i*th species inhibits the species designated modn(i+1),…, modn(i+k-1). Here, modn(*) denotes a positive residue of division of the figure (*) by *n;* if the residue equals zero, modn(*) is considered equal *n.* This system may be described as equation (1).

A similar result may be obtained for the species mutually inhibiting their reproduction at different developmental stages. For example, in the case of the pair frog–dragonfly, adult frogs eat adult dragonflies, while dragonfly larvae eat frogspawn.

Results

This work reports the results of investigating a mathematical model of hypothetical ecological networks—the model $E_1(n,k)$—in a form of autonomous system of equations:

$$\frac{dx_i}{dt} = x_i\left(\frac{\alpha}{1+\beta z_i} - x_i\right), \qquad i = 1,2,...,n, \tag{1}$$

where

$$z_1 = x_n^\gamma + x_{n-1}^\gamma + ... + x_{n-k+2}^\gamma,$$
$$z_1 = x_1^\gamma + x_n^\gamma + ... + x_{n-k+3}^\gamma,$$
$$.....................................$$
$$z_1 = x_{n-1}^\gamma + x_{n-2}^\gamma + ... + x_{n-k+1}^\gamma.$$

Here, $\alpha > 0$, $\beta > 0$, and $\gamma \geq 1$ are the parameters of the model; $k \leq n$. When the multiplier x_i is absent in the right part of equation (1), the autonomous system (1) represents the model $M_1(n,k)$ of symmetrical hypothetical gene networks (symmetrical HGN) with regulatory links of type 1 (Fadeev *et al.,* 1998; Likhoshvai *et al.,* 2003).

Evidently, all the stationary solutions for $M_1(n,k)$ are the stationary solutions for the model $E_1(n,k)$. In addition, there exist the stationery solutions of $E_1(n,k)$, among whose components there are zero solutions (additional stationary solutions). It is proved that the additional stationary solutions are Lyapunov unstable at $t > 0$. This allows the results of studying the stationary solutions of symmetrical HGN to be transferred to stationary solutions of equation (1) (Fadeev *et al.,* 1998; Likhoshvai *et al.,* 2003; Likhoshvai, Fadeev, 2003).

An efficient semi-implicit method of numerical integration of the autonomous system that takes into account the form of the right parts of system (1) is proposed.

The qualitative and numerical studies of the model $E_1(n,k)$ confirmed the statements of the known (n,k)-criterion, formulated for the model $M_1(n,k)$, when applied to the model $E_1(n,k)$: for any $\beta > 0$ there exist $\bar{\alpha} > 0$ and $\bar{\gamma} > 1$ such that at $\alpha > \bar{\alpha}$, and $\gamma > \bar{\gamma}$ the autonomous system (1), representing the model $E_1(n,k)$, has only k stable stationary solutions, if the greatest common divisor of figures n and k is equal to k. In this case, the limit cycles are absent. If $d \neq k$, the model $E_1(n,k)$ has only d stable limit cycles, whereas the stable stationary solutions are absent.

It is demonstrated numerically that the components of stable limit cycles fall into d groups; within each group, the components have the same amplitude, differing only in the phase. This allows self-oscillations of the autonomous system (1) to be described with systems of d differential equations with retarded arguments.

In particular, in the model $E_1(n,2)$ with odd n, the equation with retarded argument has the following form:

$$\frac{dz}{dt} = z\left(\frac{\alpha}{1+\beta z^{\gamma}(t-\tau)} - z\right), \qquad (2)$$

where $\tau = \frac{n-1}{2n}T$ and T is the period of the limit cycle of model $E_1(n,2)$. The necessary and sufficient conditions for the stationary solution of equation (2) to be asymptotically stable were found (Bellman, 1967). The regions of variations of the parameters α and τ where the solution of equation (2) achieves a periodic mode are determined.

An efficient algorithm for numerical integration of equation (2), used when analyzing numerically equation (2), is proposed.

Acknowledgements

The work was supported in part by the Russian Foundation for Basic Research (grants Nos. 02-04-48802, 03-04-48829, and 04-01-00458); Russian Ministry of Industry, Sciences, and Technologies (grant No. 43.106.11.0011); Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”, Program for Basic Research of the Presidium of the RAS (contract No. 10002-251/ П -25/155-270/200404-082).

References

Bellman R., Kuk K. Differential–Difference Equations. M.: Mir, 1967.

Likhoshvai V.A., Fadeev S.I. On hypothetical gene networks // Sib. Zh. Industr. Matemat. 2003. V. 4. P. 134–153.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. Problems in the theory of operation of gene networks // Sib. Zh. Industr. Matemat. 2003. V. 4. P. 64–80.

Fadeev S.I., Pokrovskaya S.A., Berezin A.Yu., Gainova I.A. The Software Package STEP for Numerical Study of Systems of Nonlinear Equations and Autonomous Systems of General Form. Novosibirsk: NGU, 1998.

METHOD FOR INTEGRATION OF DATABASES WITH COMMON SUBJECT DOMAINS

Pisarev A.*, Blagov M., Samsonova M.

Department of Computational Biology, Center for Advanced Studies, St. Petersburg State Polytechnical University, 29 Polytechnicheskaya ul., St. Petersburg, 195251, Russia
* Corresponding author: e-mail: pisarev@spbcas.ru

Keywords: *database integration, natural language interface, conceptual schema, multiagent systems*

Summary

Motivation: The integration of biological data from the heterogeneous data sources is one of the central problems of bioinformatics.

Results: We present a novel approach to the integrated retrieval of molecular biology data, based on application of the technology of multiagent systems and design of an adaptive natural language interface. Our approach allows to integrate any information resources which have a common subject domain. The architecture of the system ensures its portability across software/hardware platforms, high adaptivity to functional extensions and modifications, as well as the optimal distribution of query load between several database mirrors.

Availability: http://urchin.spbcas.ru/NLP/NLP.htm

Introduction

One of the most widespread systems of unified access to molecular biology data is SRS (Sequence Retrieval System). (Etzold *et al.*, 1996). In the 90s the language SYNTHESIS was created to develop the heterogeneous interoperable environments of information resources (Kalinichenko *et al.*, 2000). The semantic integration of information resources is provided with the use of broker mechanisms and ontological specifications. As an example of implementation of this approach an objective broker for integration of information resources in the area of molecular genetics is currently being developed in the Institute of Cytology and Genetics of the SB RAS [http://wwwmgs.bionet.nsc.ru/mgs/systems/genenet/]. A similar approach to the integration of molecular biology databases is applied in the European Bioinformatics Institute (Coupaye, 1999). The essential limitation of the approaches mentioned above is that they require using and supporting the actuality of definitions in the specialized languages if the model of subject domain is changed. In this paper we present a new method for integration of databases with a common subject domain. The key components of this method are the conceptual schema of knowledge domain and domain oriented dictionaries; processor of natural language (NL) queries to a database; multiagent architectute to integrate the results of information retrieval from different databases. This method was applied to integrate the information about expression of segmentation genes in fruit fly *Drosophila.*

Material and Methods

Databases. The genetic network which controls segmentation is one of a few genetic networks fully characterized at genetic and functional level (Ingham, 1988). The initial determination of segments is the consequence of expression of 16 genes. Spatio-temporal expression of segmentation genes is studied in details. The expression of most of segmentation genes starts at embryo stage 4 when the blastoderm is syncytial and not divided into cells and continues at embryonic stage 5. At this time the segments are determined and invagination of membranes and cell cellularization happens. Besides expression at the time of segment determination most segmentation genes are active at later developmental times, and some of them are stably expressed throughout the life of the fly. Segmentation genes control muscule formation, neurogenesis and other developmental processes (Campos-Ortega, Hartenstein, 1985).

Currently the information about expression of segmentation genes is stored in several databases, namely FlyEx (http://urchin.spbcas.ru/flyex) and FlyEx mirror at University of New York at Stony Brook (http://flyex.ams.sunysb.edu/flyex), Mooshka (http://urchin.spbcas.ru/Mooshka), FlyBase ((http://flybase.bio.indiana.edu/) and *In situ* Database (http://www.fruitfly.org/cgi-bin/ex/insitu.pl). FlyEx and Mooshka store the information about expression of segmentation genes at the time of segment determination. FlyBase contains the information about time and place of expression of each segmentation gene through the life of the fly. *In situ* database, which is being developmed now, stores the images of expression patterns of segmentation genes at all points of development. The determination of segments is the subject of intensive research over the last two decades and the integration of data on segmentation gene expression from the heterogeneous data sources mentioned above is an essential part of the work of researchers in both industry and academia. As the integration-by-navigation is a tedious process, we have developed the system for automatic integration of this information.

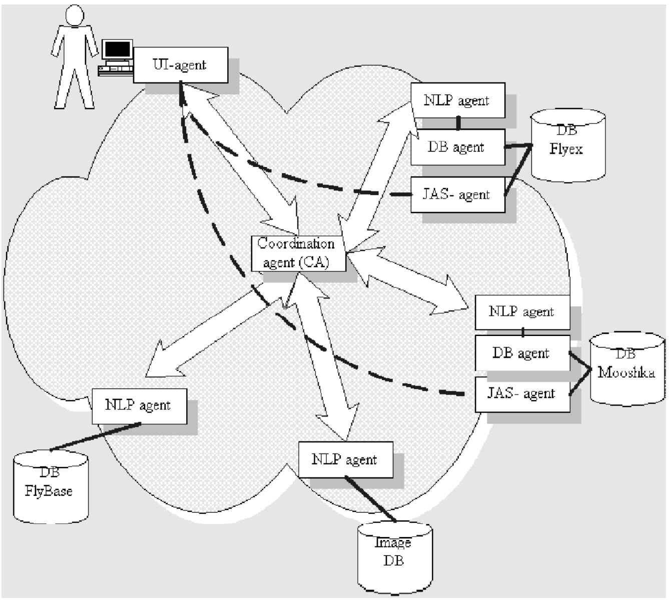


Fig. The architecture of the system.

System architecture. Our goal was the development of technology for integration of databases, which will support the use of traditional Web browsers for information retrieval; continuous work, when new databases are added or old one are removed; failure-resistance, if the malfunction of hardware or software components happens; the optimal distribution of queries to increase the performance of the system. To meet this requirements we have developed the multiagent system, which conponents are user interface agents, coordinating agent, NLP agents, DB agents and JAS agents, which serve to visualize data in different formats (Fig.). The coordinating agent receives a request from a user interface agent and distributes this request among NLP agents. A NLP agent translates a query in NL in SQL and sends it to the DB agent, which in turn requests the database and transfers the result of the query back to the NLP agent. The later formats the result and sends it to the coordinating agent. Coordinating agent, if necessary, fuses results selected from different databases and transmits the resultant information to the user interface agent.

There exist many standard protocols of agent's interaction. Most widely used are ACL (KQML/KIF) [FIPA], CORBA [ODMG], Java RMI (Gavrilova, Horoshevsky, 2000). Till recently these standards imply direct connection between agents, what was a major hindrance to the integration of information resources with FireWall and Proxy servers. We have solved this problem by developing agents interacting via the HTTP protocol and implementing all agents in Java.

The distinctive feature of the approach used in this work is the development of the algorithm of self-organization, which ensures dynamic reconfiguration of the system and optimal distribution of queries with regard to its real load.

NLP agent. We use the language understanding technology based on semantic approach. This technology interprets the grammatical and lexical units of any NL into concepts of a knowledge domain (Samsonova *et al*., 2003). These concepts are introduced in conceptual schema. The conceptual schema of the information on expression of segmentation genes in fruit fly *Drosophila* is an oriented graph, which nodes are concepts of the knowledge domain, and edges define relations between the concepts. This schema serves as a connecting link between the text of a query and a database schema. It helps the natural language processor to interpret higher or lower level concepts and synonyms, as well as equivocal and jargon terms. Moreover, making the specifications of domain knowledge explicit, the conceptual schema guides a user to learn the meaning of each term.

The procedure of processing of NL queries transforms different combinations of word forms to a limited set of terms of the logical level, which are used for a generation of the SQL queries to a database. Firstly, synonyms and high level concepts are substituted by terms of logical level, which are mapped on the database objects. At the step called 'Search in the Dictionary' an initial chain of semantic components is constructed. The step 'Semantic analysis' converts the initial chain of semantic components into a semantic network, which formally represents a query. At the 'SQL query generation and optimization' step the semantic network is converted into the SQL query to the database. The 'Advanced processing of queries' step performs the subject domain specific processing of queries, e.g., displays data in different views (as a table, graph or image).

Implementation

Information retrieval. To formulate and execute queries to the system the HTML form 'Natural Language Interface' is to be filled by a user. The text of a query is entered in the text field QUERY; the checkboxes allow to define databases used to retrieve information. The list QUERY EXAMPLES contains a set of predefined standard queries for convenience. By pressing the button SEND QUERY a query will be executed and after a while a result of the query will appear in a new browser window. In the upper part of this window the query in natural language is displayed, in which words used to retrieve the information from databases are shown in red. Below the query the result of retrieval of information from each database is displayed as a table. The SQL query, automatically generated by the system, is presented below the result. SQL query can be edited and returned to the server by pressing the button SEND QUERY.

Selection of the link SWITCH TO RUSSIAN calls the Russian version of the query form. The queries in Russian are submitted and executed similarly to queries in English.

The capabilities of the NL processor. To formulate a NL query a user can use any concept described in the conceptual scheme, type the words in any word form, formulate a query as a whole phrase or as a list of keywords, use synonyms or even laboratory jargon, combine selection criteria in a query using logical operators AND, OR, NOT. The query *'How many ...?'* allows to count rows satisfying any criterion (e.g., *'How many embryos are scanned for expression of bcd and belong to late temporal classes?'*; the query *'Display pattern ...'* returns a pattern of segmentation gene expression.

The following combinations of semantic constructions are supported: *larger than, greater than, more than, >, >=, <=, <, less than, smaller than, from n to m, n - m.*

Conclusions

Our approach allows to integrate any information resources (published in the Internet, as well as stored locally) which have a common subject domain. Its benefits are in possibility to formulate arbitrary queries in various languages (in English and in Russian, currently), the optimal transformation of queries from natural language to SQL, as well as in opportunity to present information visually as hyperschemata. Other advantages are the simplicity in access to information and integration of new databases, adaptivity with respect to changes in a knowledge domain and user's views, increase of the robustness of the system as well as the optimization of distribution of queries load between several database mirrors.

Acknowledgements

The support of this work by EC grant IST-1999-11009 and GAP award RMO-1323 is gratefully acknowledged.

References

Campos-Ortega J.A., Hartenstein V. The Embryonic Development of *Drosophila melanogaster*. Springer-Verlag: Berlin. 1985.

Coupaye T. Wrapping SRS with CORBA: from textual data to distributed objects // Bioinformatics. 1999. V. 15(4). P. 333–338.

Etzold T., Ulyanov A., Argos P. SRS: Information Retrieval System for Molecular Biology Data Banks // Methods in Enzymolog. 1996. 226.

Gavrilova T.N., Horoshevsky V.F. Knowlegde Bases for Intelligent Systems. St.Petersburg, Piter. 2000.

Ingham P.W. The molecular genetics of embryonic pattern formation in Drosophila // Nature. 1988. V. 335. P. 25–34.

Kalinichenko L.A., Briukhov D.O., Skvortsov N.A., Zakharov V.N. Infrastructure of the subject mediating environment aiming at semantic interoperability of heterogeneous digital library collections. Institute for Problems of Informatics RAS // Proc. of the Second All Russian Conference on Digital Libraries. Protvino, September, 2000.

Samsonova M., Pisarev A., Blagov M. Processing of natural language queries to a relational database // Bioinformatics. 2003. V. 19. Suppl. 1. i241–i249.

QUANTITATIVE APPROACH TO THE FUNCTIONAL GENOMICS OF DEVELOPMENT

Samsonova M. [1]*, ***Surkova S.***[1], ***Jaeger J.***[2], ***Reinitz J.***[2]

[1] Department of Computational Biology, Center for Advanced Studies, St. Petersburg State Polytechnical University, 29 Polytechnicheskaya ul., St. Petersburg, 195251 Russia; [2] Department of Applied Mathematics and Statistics, and Center for Developmental Genetics, Stony Brook University, Stony Brook, NY 11794-3600, USA

* Corresponding author: e-mail: samson@spbcas.ru

Keywords: *Drosophila, segmentation, gene circuits, gene expression, quantitative data, morphogens*

Summary

Motivation: The ability to quantify molecules and events in living cells is essential to understand how a biological system functions. This ability is especially important in study of development at early stages, at which the events in determination and pattern formation take place. Quantitative measurements in functional genomics of development become feasible due to the advent of fluorescent protein technologies and sophisticated light microscopy techniques.

Results: In this paper we will demonstrate the power of quantitative approach by presenting the results of analysis of a particular developmental system, namely segment determination in the fruit fly *Drosophila melanogaster*. We show that with quantitative information we can address questions which had never been addressed before. In particular, a quantitative approach allows us to describe the segment determination system at a fine level of detail, estimate the precision of its functioning, as well as to explain and predict the system's behaviour.

Availability: All data are available on request.

Introduction

In early *Drosophila* embryogenesis the segmented body plan is established through a cascade of maternally and zygotically expressed segmentation genes. The zygotic genes have been classified according to their mutant phenotypes and expression patterns. Gap genes are expressed in one to three broad domains, pair-rule genes form seven transverse stripes and segment polarity genes manifest in patterns of fourteen stripes.

It is widely believed that maternal gradients directly define the territories of zygotic gap and pair-rule gene expression, which then convey the positional information encoded by these gradients to the segmental prepattern formed by the segment-polarity genes *engrailed (en)* and *wingless (wg)* (Akam, 1987; Ingham, 1988; Reinitz *et al*., 1998). However, recent results suggest that the maternal gradient of Bicoid (Bcd) requires synergism with maternal Hunchback (Reinitz, Sharp, 1995; Simpson-Brose *et al*., 1994) and cannot account for the precise positioning of zygotic *hb (hunchback)* expression by itself (Houchmandzadeh *et al*., 2002).

Material and Methods

Images of gene expression patterns were acquired as described in (Kosman *et al*., 1997). These images serve as raw material for the quantification of gene expression. The quantitative data were obtained in several steps. For each step, a specialized method for image or data processing was developed and implemented (Kozlov *et al*., 2000; Myasnikova *et al*., 1999; 2001). Standard statistical procedures (StatSoft Statistica package) were used to validate the significance of temporal changes in the localization of expression domains. The positional variability of expression patterns was estimated by computing the standard deviations of positions of characteristic features. We use gene circuits, a data-driven mathematical modeling method to reveal the mechanism for gap domain shifts (Reinitz, Sharp, 1995). Model output was compared to expression data for all gene products at all time points where data was available and the sum of squared differences between model and data was minimized using Parallel Lam Simulated Annealing (Chu *et al.*, 1999).

Dataset. We have obtained quantitative data on the expression of 14 segmentation genes in 954 individual embryos belonging to cleavage cycles from 11 to 14A. These data were used to construct reference data for most segmentation genes at 9 time points (Kozlov *et al.*, 2002; Poustelnikova *et al.*, 2004). The reference data demonstrate the typical features of expression of each segmentation gene at a given developmental time and with resolution to a single nucleus.

Implementation and Results

Quantitative analysis of the dynamics of formation of segmentation gene expression domains. Most expression domains of gap and pair-rule genes do not form in one place but change their position with time. Central and posterior domains of gap genes shift anteriorly, while anterior gap domains move in the opposite direction; all stripes of pair-rule genes except anterior ones move in the anterior direction. These domain shifts are of the size of a pair-rule stripe at the end of cleavage cycle 14A and thus are very important for the positioning of domains of downstream genes.

Estimation of the precision of development. The maternal gradients of Bcd and Caudal (Cad) cannot account by themselves for positioning of downstream segmentation genes: these gradients have very high spatial variability, whereas the expression domains of gap and pair-rule gens are positioned with precision of about one nucleus.

The mechanism of gap domain shifts in the trunk region of an embryo. Here we address this question by modelling the behaviour of gap gene network including *bcd*, *cad*, *hb*, *Kr (Kruppel)*, knirps (*kni)*, giant (*gt)* and *tailless (tll)* in the region of the embryo between 32 and 92 % anteroposterior (AP) position (where 0 % is the anterior pole) during cleavage cycles 13 and 14A. Gap gene circuits are able to reproduce observed gap gene expression patterns with high precision and temporal resolution and faithfully reproduce expression domain shifts of the central domain of *Kr;* as well as the posterior domains of *kni* and *gt* (Jaeger *et al.*, 2004). Dynamical shifts of gap gene expression domains are reflected at the level of the rate of change in protein concentration: in each of *Kr*, *kni* and *gt* expression domains protein synthesis dominates anteriorly, protein decay posteriorly. The combination of anterior synthesis and posterior decay leads to the anterior shift of the expression domain. The analysis of shifts of gap domain boundaries has shown that the posterior boundaries of the trunk gap domains shift due to dynamical mechanisms based on asymmetric repressive regulatory interactions between neighbouring gap genes. Shifts of anterior gap domain boundaries can be considered secondary effects of the dynamic behaviour of posterior boundaries, as they either follow posterior boundary shifts of more anterior gap genes or are due to the sharpening of the posterior boundaries of anterior *gt* and *hb* domains.

Conclusions

By quantifying the expression of segmentation genes with resolution to a single nucleus we were able to describe the behaviour of this system at a fine level of detail, estimate the precision of positioning of segmentation gene expression domains, as well as to explain the mechanism of domain shifts in the trunk region of the embryo.

We have found that the posterior gap domains shift because of the regulative interactions between neighbouring gap genes. The absence of a clear hierarchy among these regulatory interactions implies an emergent rather than hierarchical topology of the underlying gene network (Salazar-Cuidad *et al.*, 2000). Thus the maternal gradients of Bcd, Hb and Cad are not sufficient for the positioning of gap gene domains and hence do not qualify as morphogens *sensu strictus*. An active role of target tissue in specifying positional information contradicts the traditional distinction between the instructive role of maternal morphogens and their passive interpretation (Wolpert, 1969). Moreover, the dynamical nature of positional information as encoded by expression boundaries in the blastoderm suggests that positional information in the blastoderm embryo can no longer be seen as a static coordinate system imposed on the embryo by maternal morphogens (Wolpert, 1969). Rather, it needs to be understood as a dynamic process underlying positioning of expression domain boundaries based on both external inputs by morphogens and tissue-internal feedback among target genes.

Acknowledgements

This work was supported by NIH Grants RR07801 and TW01147, and by CRDF GAP Award RBO-1286.

References

Akam M. The molecular basis for metameric pattern in the *Drosophila* embryo // Development. 1987. V. 101. P. 1–22.

Chu K.W., Deng Y., Reinitz J. Parallel simulated annealing by mixing of states // J. Comp. Phys. 1999. V. 148. P. 646–662.

Houchmandzadeh B., Wieschaus E., Leibler S. Establishment of developmental precision and proportions in the early *Drosophila* embryo // Nature. 2002. V. 415. P. 798–802.

Ingham P.W. The molecular genetics of embryonic pattern formation in *Drosophila* // Nature. 1988. V. 335. P. 25–34.

Jaeger J., Blagov M., Kosman D., Kozlov K.N., Manu, Myasnikova E., Vanario-Alonso C.E., Samsonova M., Sharp D.H., Reinitz J. Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster* // Genetics. 2004, in press.

Myasnikova E., Kosman D., Reinitz J., Samsonova M. Spatio-temporal registration of the expression patterns of *Drosophila* segmentation genes // Proc. 7th Intl. Conf. on Intel. Syst. Mol. Biol. 1999. Menlo Park, California, AAAI Press. P. 195–201.

Myasnikova E., Samsonova A., Kozlov K., Samsonova M., Reinitz J. Registration of the expression patterns of Drosophila segmentation genes by two independent methods // Bioinformatics. 2001. V. 17. P. 3–12.

Kosman D., Reinitz J., Sharp D. Automated assay of gene expression at cellular resolution // Pac. Symp. on Biocomput. 1997. P. 6–17.

Kozlov K., Myasnikova E., Samsonova M., Reinitz J., Kosman D. Method for spatial registration of the expression patterns of Drosophila segmentation genes using wavelets // Computational Technologies. 2000. V. 5. P. 112–119.

Kozlov K., Myasnikova E., Pisarev A., Samsonova M., Reinitz J. A method for two-dimensional registration and construction of the two-dimensional atlas of gene expression patterns *in situ* // In Silico Biol. 2002. V. 2. P. 125–141.

Poustelnikova E., Pisarev A., Blagov M., Samsonova M., Reinitz J. A database for management of gene expression data *in situ* // Bioinformatics. 2004, in press.

Reinitz J., Sharp D.H. Mechanism of *eve* stripe formation // Mech. Dev. 1995. V. 49. P. 133–158.

Reinitz J., Kosman D., Vanario-Alonso C. E., Sharp D.H. Stripe forming architecture of the gap gene system // Dev. Gen. 1998. V. 23. P. 11–27.

Salazar-Ciudad I., García-Fernandez J., Solé R.V. Gene networks capable of pattern formation: from induction to reaction-diffusion // J. Theor. Biol. 2000. V. 205. P. 587–603.

Simpson-Brose M., Treisman J., Desplan C. Synergy between the Hunchback and Bicoid morphogens is required for anterior patterning in D*rosophila* // Cell. 1994. V. 78. P. 855–865.

Wolpert L. Positional information and the spatial pattern of cellular differentiation // J. Theor. Biol. 1969. V. 25. P. 1–47.

A SIMULATION MODEL OF EnvZ-OmpR TWO COMPONENT SYSTEM IN *ESCHERICHIA COLI*

Srividhya K.V., Krishnaswamy S.*

Bioinformatics Centre, School of Biotechnology, Madurai Kamaraj University, Madurai 625 021, Tamil Nadu, India
* Corresponding author: e-mail: mkukrishna@rediffmail.com

Keywords: *two component systems, osmoregulation, Escherichia coli, E-CELL, simulation, sensor kinase, response regulators*

Summary

Motivation: The response of *Escherichia coli* to high and low osmolyte concentration in the environment is controlled by the EnvZ–OmpR two component signal transduction system. High osmolarity leads to increase in the levels of the outer membrane porin OmpC. Low osmolarity results in decrease of OmpC and increase of OmpF levels. A quantitative simulation model will be useful in understanding the controlled expression of OmpC and OmpF porins in response to the osmolyte concentration.

Results: A quantitative model of EnvZ-OmpR osmoregulatory switch in *Escherichia coli* was constructed by integrating a total of 28 enzyme rate equations using the E-CELL system with available experimental data from literature. Changes in volume, ATP, EnvZ and OmpR did not alter the relative porin production, at low and high osmolarity conditions, highlighting the robust nature of the system.

Introduction

Two-component regulatory systems are widely used signaling machinery of bacterial adaptive responses comprising of sensor kinases and response regulators (Stock *et al*., 1989) .The first component sensor-transmitter spanning the cytoplasmic membrane, has sensory domain and a transmitter domain with specific Histidine (His) residue, which is autophosphorylated utilizing ATP at this site. The receiver-regulator, localized cytoplasmically, perceives the environmental signal through phospho transfer from Sensor His (Alex, Simon, 1999) to Aspartate (Asp) residue and regulates gene expression at DNA level (Mattison *et al*., 2002). Computer simulations of chemotaxis two component system has been extensively studied (Bourret *et al*., 1989). Osmoregulation in *Escherichia coli*, comprises of EnvZ sensor and OmpR regulator (Csonka, 1991) controlling the expression of the major outer membrane proteins, OmpC and OmpF for diffusion of hydrophilic molecules across the membrane (Mizuno *et al*., 1983). The osmotic signal modulates the ratio of the kinase to phosphatase activity of EnvZ (Yang *et al*., 1993). At low osmolarity Phosphatase activity of EnvZ predominates the kinase activity resulting OmpF expression. At high osmolarity kinase activity of EnvZ is trigerred resulting in repression of *ompF* gene expression (Bergstrom *et al*., 1998; Head *et al*., 1998) and OmpC expression (Bergstrom *et al*., 1998).

Methods

The E-CELL Windows version 2.25 was employed for simulation (www.e-cell.org). along with the third party software namely Active Perl, JRE (java runtime environment) and Borland C++ compiler (http://www.borland.com) essential for running simulations. The computational model of osmoregulatory switch is based on the mathematical model (Batchelor, Goulian, 2003). The information defining all the components of the osmoregulatory switch including reactions, rate constants, kinetics, simulation time and time interval were incorporated in the rule and script files.

Results and Discussion

The present model is built on the assumption of the *in vivo* condition considering *Escherichia coli* cells grown in mid–log phase with OmpR and EnvZ levels of 3500 and 100 molecules in cell respectively. The high osmolarity medium has additional sucrose (NB (Nutrient Broth) +20 % sucrose) (Lilijestrom *et al.*, 1988). The ratio of OmpR to EnvZ is reported to be constant, assuming the cell volume to be 10^{-15} liters (Wanner, 1996). Phosphorylation of only 3.5 % of total OmpR molecules in a cell (2024 molecules) would be enough to activate OmpF expression at low osmolarity whereas at high osmolarity the phosphorylation of about 10 % of total cellular OmpR molecules (3500 molecules per cell) is needed to promote the expression of OmpC and to repress the expression of OmpF (Frost *et al.*, 1990).

Simulation of Low osmolarity and High osmolarity

Low osmolarity conditions were simulated assuming that *Escherichia coli* cells are grown in normal nutrient medium. The sucrose levels were maintained as normal (around 150 molecules) (Cai, Inouye, 2002). OmpF synthesis is seen to be triggered at the start of simulation. The entire trend of OmpF synthesis – gradual increase, steady state and final saturation are represented graphically (Fig. 1). For high osmolarity conditions, the model generated incorporates the concentration of sucrose with the assumption that *Escherichia coli* cells are grown in nutrient broth with 20 % additional concentration of sucrose (1.11M equivalent) (Cai, Inouye, 2002). With this injection stimulus, the kinase activity of EnvZ is enhanced leading to OmpC expression (Fig. 2).

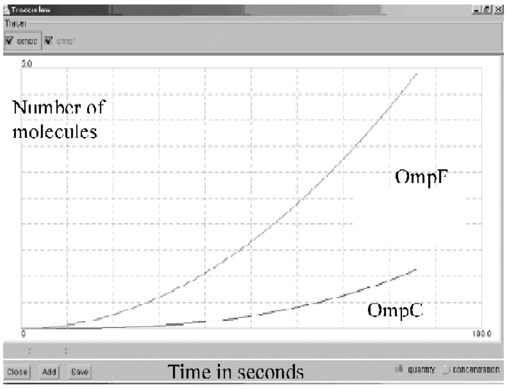


Fig. 1. Tracer window display of porins at Low osmolarity conditions.

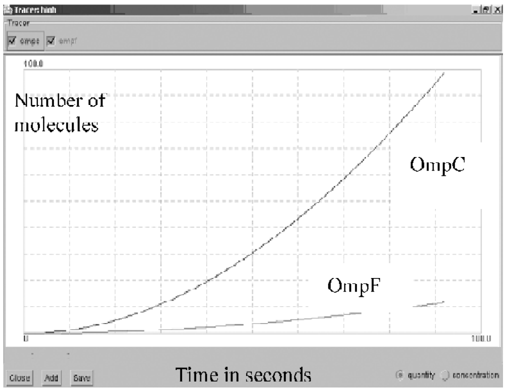


Fig. 2. Tracer window display of porins at High osmolarity conditions.

Effect of Volume changes ATP, Sensor, and Regulator levels over porin production

Shrinkage of *Escherichia coli* is associated with osmotic challenge (Koch, 1984). On incorporating a volume decrease of 10 % and 20 % from the specified 10^{-15} L, the simulation did not affect the relative level of porin production although the reach of saturation was rapid (Table). An ATP level of 3 and 5mM has been reported in exponentially growing *Escherichia coli* cells (Koebmann *et al.*, 2002).

The earlier report on ATP increase leading to plasmolysis thereby leading to crowding of molecules (Wood, 1999) and increasing levels of EnvZ and OmpR till 10 fold (1000 molecules), does not seem to affect the ratio of porins in the simulated system highlighting robust nature of the switch. As per the mathematical model. EnvZ is required for the maximal OmpC production and for efficient induction of OmpC at high osmolarity (Frost *et al.*, 1988). This hypothesis is cross verified with the simulation model.

Table. Effect of ATP, EnvZ, OmpR and volume at high and low osmolarity

Condition	Low Osmolarity				High Osmolarity			
	Saturation time OmpF(s)	OmpF	OmpC	Ratio OmpF/OmpC	Saturation time OmpC(s)	OmpC	OmpF	Ratio OmpC/OmpF
Low ATP	30	244	100	2.4	18	557	129	4.3
High ATP	26	242	106	2.2	13	557	130	4.2
Low EnvZ	120	246	6	41	85	550	246	2.2
High EnvZ	135	244	104	2.3	80	550	125	4.4
Low OmpR	125	35	13	2.6	70	52	14	3.7
High OmpR	130	2176	889	2.4	115	5309	1210	4.3
Low Volume	150	245	108	2.2	73	537	124	4.3
High Volume	130	243	100	2.4	80	548	125	4.3

Conclusion

We have shown porin regulation at high and low osmolyte concentrations mediated through EnvZ. The preliminary simulation experiment indicates that both reaching steady state expression and saturation is delayed in the case of OmpC compared to OmpF. The relative porin production seem to be unaltered with changes in cell volume changes, ATP, EnvZ and OmpR at low and high osmolarity conditions. But the reach of saturation was rapid at high and low osmolarity with altered levels of the above components. Though the model simulated here is a simplified description of the EnvZ/OmpR system, experimental analysis will help improve the model. The model is a reasonable starting point for building sophisticated models and explaining quantitative features of the system.

Acknowledgements

We acknowledge the use of Bioinformatics centre facility funded by DBT, Govt of India and CSIR for project support under the NMITLI programme.

References

Alex L.A., Simon M.I. Protein Histidine kinases and signal transduction in prokaryotes and eukaryotes // Trends.Genet. 1994. V. 10. P. 133–139.

Batchelor E., Goulian M. Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system // Proc. Natl Acad. Sci. 2003. V. 100. P. 691–694.

Bergstrom L.C., Qin L., Harlocker S.L., Egger L.A., Inouye M. Hierarchical and co-operative binding of OmpR to a fusion construct containing the ompC and ompF upstream regulatory sequences of *Escherichia coli* // Genes to Cells. 1998. V. 3. P. 777–788.

Bourret R.B., Hess J.F., Borkovich K.A., Pakula A.A., Simon M.I. Protein phosphorylation in chemotaxis and two-component regulatory systems of bacteria // J. Biol. Chem. 1989. V. 264. P. 7085–7088.

Cai S.J., Inouye M. EnvZ-OmpR Interaction and Osmoregulation in *Escherichia coli* // J. Biol. Chem. 2002. V. 277. P. 24156–24161.

Csonka L.N., Hanson A.D. Prokaryotic osmoregulation // Annu. Rev. Microbiol. 1991. V. 45. P. 569–606.

Frost S., Delgado J., Ramakrishnan G., Inouye M. Regulation of OmpC and OmpF expression in *Escherichia coli* in the absence of EnvZ // J. Bact. 1988. V. 170. P. 5080–5085.

Frost S., Delgado J., Rampersaud A., Inouye M. *In vivo* phosphorylation of OmpR, the transcription activator of the *ompF* and *ompC* genes in the *Escherichia coli* // J. Bacteriol. 1990. V. 172. P. 3473–3477.

Head C.G., Tardy A., Kenny L.J. Relative binding affinities of OmpR and OmpR-Phosphate at the *ompF* and *ompC* regulatory sites // J. Mol. Biol. 1998. V. 281. P. 857–870. http://www.borland.com

Koch A.L. shrinkage of growing *Escherichia coli* cells by osmotic challenge // J. Bacteriol. 1984. V. 159. P. 1979–1984.

Koebmann B., Westerhoff H.V., Snoep J.L., Nilsson D., Jensen P.R. The glycolytic Flux in *Escherichia coli* is controlled by the Demand for ATP // J. Bacteriol. 2002. V. 184. P. 3909–3916.

Lilijestrom P., Laamanen I., Palva E. The EnvZ protein of *Salmonella typhimurium* LT-2 and *Escherichia coli* K-12 is located in the cytoplasmic membrane // FEMS Mirobiol. Letters. 1988. V. 36. P. 145–150.

Mattison K., Rand O., Kenney L.J. The linker region plays an important role in the interdomain communication of the resposne regulator OmpR // J. Bact. 2002. V. 277. P. 32714–32721.

Mizuno T., Chou M.Y., Inouye M. A comparative study on the genes for three porins of the *Escherichia coli* outer membrane: DNA sequence of the osmoregulated *ompC* gene // J. Biol. Chem. 1983. V. 258. P. 6932–6940.

Stock J.B., Ninfa A.J., Stock A.M. Protein phosphorylation and regulation of adaptive responses in bacteria // Microbiol Rev. 1989. V. 53. P. 450–490.

Wanner B.L. In *Escherichia coli* and Salmonella / Ed. F.C. Beidhardt. American Society of Microbiology, Washington. D.C., 1996. V. 1. P. 1359.

Wood J.M. Osmosensing by bacteria: signals and membrane – based sensors // Microbiol. and Mol. Biol. Rev. 1999. V. 63. P. 230–262. www.e-cell.org

Yang Y., Park H., Inoyue M. Requirement of both kinase and phosphatase activities of *Escherichia coli* receptor taz1 for ligand-dependent signal transduction // J. Mol. Biol. 1993. V. 231. P. 335–342.

TEMPORAL AND SPATIAL PRECISION IN FORMATION OF SEGMENTATION GENE EXPRESSION DOMAINS IN *DROSOPHILA*

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

St.Petersburg State Polytechnical University, Russia
* Corresponding author: e-mail: surkova@spbcas.ru

Keywords: *Drosophila , segmentation, gene expression, positional information, temporal precision*

Resume

Motivation: Over the last two decades the mechanism of segment determination in early *Drosophila* embryo is a subject of intensive research. We investigate the mechanisms of segment determination by the integrated program of mathematical modeling and experiment. Estimation of spatial and temporal variability of segmentation gene expression patterns is necessary to validate the model of segment determination introduced in (Reinitz, Sharp, 1995).

Results: Spatial variability of Bicoid and Caudal gradients is several times higher than the variability in positioning of zygotic segmentation genes. However, the variability of the early domain of *even-skipped* expression is comparable to that of maternal gradients. Each pair-rule stripe is formed with a distinctive temporal precision. Temporal variability of stripe formation is different for different stripes of one pair-rule gene.

Availability: All data is available from authors.

Introduction

In early *Drosophila* embryogenesis the segmented body plan is established through a cascade of maternally and zygotically expressed segmentation genes. At the first step maternal genes set the anteroposterior (A-P) polarity of the egg and form the anterior and posterior protein gradients. These gradients are responsible for diverse events cued at different points along the egg axis in a concentration-dependent fashion through direct effects on downstream zygotic genes, either activating or repressing them (reviewed in Akam, 1987). The zygotic genes have been classified according to their mutant phenotypes and expression patterns. ‘Gap’ genes are expressed in one to three broad domains, ‘pair-rule’ genes form seven transverse stripes and ‘segment polarity’ genes manifest in patterns of fourteen stripes about one-cell wide. The translation of information stored in maternal gradients into pair-rule and segment-polarity regular striped patterns is a challenge, which draws the attention of many research groups during the last two decades. An important step towards the solution of this problem is the precise characterization of dynamics of segmentation gene expression.

The significant biological question is the level of spatial and temporal variability of segmentation gene expression patterns in individual embryos. The measurement of this level gives us a unique possibility to estimate the precision of development.

We investigate the mechanisms of segment determination by the integrated program of mathematical modeling and experiment. Experimental work is performed to acquire data on segmentation gene expression at cellular resolution (Kozlov *et al*., 2000, Myasnikova *et al*., 2001), while mathematical modeling is based on method known as gene circuits (Reinitz, Sharp, 1995; Jaeger *et al*., 2004).

Materials and Methods

Images of gene expression patterns were obtained 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 809 wild type embryos from cleavage cycle 14A (~50 minutes long) and 93 embryos from cycle 13 (~20 minutes long). The embryos were scanned for the expression of 11 segmentation genes. The images of embryos from cycle 14A were distributed by visual inspection of pair-rule gene expression patterns into 8 temporal equivalence classes (Myasnikova *et al.*, 2001). In this study we used 1D data in the 10 % central strip along the A-P axis of an embryo. As the characteristic features of segmentation gene expression domains we considered the A-P positions of peaks and points, where the level of gene expression reached the predefined threshold. The positions of expression maxima for pair-rule stripes were extracted by means of the wavelet decomposition of the signal (Kozlov *et al.*, 2000), while patterns of gap genes and maternal gradients were approximated by quadratic splines (Myasnikova *et al.*, 2001). The positional variability of expression patterns was estimated by computing the standard deviations of positions of characteristic features using StatSoft Statistica package.

Results and Discussion

Spatial precision in formation of segmentation gene expression domains. The analysis of spatial variability of segmentation gene expression patterns can serve as a framework to estimate the precision of pattern formation. The first step of pattern formation in *Drosophila* is the establishment of the A-P polarity of the egg and the formation of the anterior and posterior gradients of proteins encoded by the maternal genes (Fig. 1A). Positional information in these gradients is transmitted to the downstream gap genes. Recently Houchmandzadeh *et al.* (2002) had shown that Bcd gradient shows high embryo-to-embryo spatial variability and that this variability is strongly decreased at the level of expression of downstream gap gene *hb*.

In our work we have confirmed these results and extended our analysis to other segmentation genes.

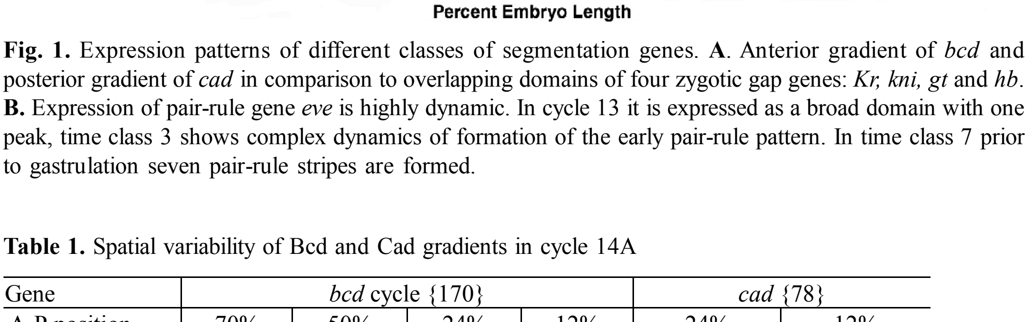


Fig. 1. Expression patterns of different classes of segmentation genes. **A**. Anterior gradient of *bcd* and posterior gradient of *cad* in comparison to overlapping domains of four zygotic gap genes: *Kr, kni, gt* and *hb*. **B.** Expression of pair-rule gene *eve* is highly dynamic. In cycle 13 it is expressed as a broad domain with one peak, time class 3 shows complex dynamics of formation of the early pair-rule pattern. In time class 7 prior to gastrulation seven pair-rule stripes are formed.

Table 1. Spatial variability of Bcd and Cad gradients in cycle 14A

Gene	*bcd* cycle {170}				*cad* {78}	
A-P position	70%	50%	24%	12%	24%	12%
St. Dev.	**3.52**	**4.37**	**5.64**	**5.76**	**6.57**	**4.41**

Table 1. *A-P position* stands for the positions of points where gradients cross the thresholds of 70, 50, 24 and 12 % of maximal fluorescence intensity.

Table 2. Spatial variability of positioning of gap domains at temporal class 5

Domain	*gt* ant {20}		*hb* ant {29}	*Kr* cent {46}	
Boundary	A	P	P	A	P
St. Dev.	**1.42**	**1.53**	**1.26**	**1.08**	**1.04**

Domain	*kni* post{22}		*gt* post {23}		*hb* post {30}	
Boundary	A	P	A	P	A	P
St. Dev.	**1.0**	**1.18**	**0.97**	**1.48**	**1.06**	**1.46**

Table 3. Spatial variability of *bcd, hb, Kr* and *eve* patterns at cycle 13 and early cycle 14A

Gene, Dev. Time	*bcd* cycle 13 {80}				*hb* cycle 13 {25}	*Kr* cycle 13 {20}
A-P position	70%	50%	24%	12%	P	Peak
St. Dev.	**3.14**	**4.36**	**4.02**	**3.80**	**2.27**	**2.47**

Gene, Dev. Time	*eve* cycle 13 {75}			*eve* time class 1 {78}			*eve* time class 2{63}	
A-P position	A	peak	P	A	peak	P	A	peak
St. Dev.	**2.38**	**2.39**	**3.65**	**2.07**	**2.20**	**3.42**	**1.38**	**1.65**

In Tables 1, 2 and 3: Sample sizes are shown in braces. *St.Dev.* stands for a standard deviation, which is a measure of noise in positioning expression domains. *A* and *P* stand for anterior and posterior boundaries of the domains, i.e. points corresponding to 50 % concentration thresholds.

We have demonstrated that the positional variability of Bcd gradient is in the range from 3 to 6 embryo nuclei both at cycles 13 and 14A, and that at cycle 14A Cad gradient has similar variability in the range from 6 to 4 nuclei (Tables 1, 3).

In contrast, the positioning of gap expression domains is less variable. The positional error in specification of expression domains of *hb* and *Kr* ranges from 2.0 to 2,5 % EL at cycle 13 and further declines by about one half at cycle 14A. At this cycle other gap domains (Table 2) and pair-rule stripes (not shown) are positioned with a precision of about one nucleus.

The posterior boundary of *eve* early expression domain, which is the exact region where future stripes will later arise (Fig. 1B), has the spatial variability similar to that of the maternal gradients both at cycle 13 and time class 1 of cycle 14A (Table 3). However, the positional error of *eve* in this region declines by 2–3 times starting from time class 2 when *eve* stripes begin to form. This fact supports the previously proposed hypothesis that the early *eve* domain plays a different role in the regulation of downstream genes than the late pattern consisting of seven stripes. In summary, our results demonstrate that the noise in positional information inherent to the maternal gradients and gentle posterior boundary of the *eve* early expression domain is decreased by several times at the level of gap and pair-rule gene expression.

Temporal precision of stripe formation. Formation of seven stripes of pair-rule genes proceeds through a set of early transient patterns (Fig. 1B). The shape of late pair-rule stripes is almost similar in all pair-rule genes, however, all these domains are formed in a different way, at different time and with a different temporal precision.

The quantitative data on gene expression allows us to estimate the temporal variability of one

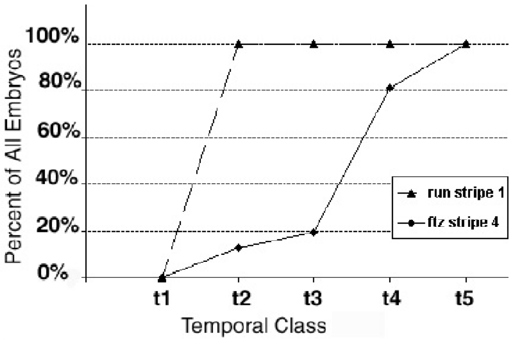


Fig. 2. Temporal variability in formation of pair-rule stripes.

stripe formation with a precision of about 6 minutes as this is the temporal resolution of our dataset. We have found that a temporal precision of stripe formation can vary in a wide range from less than 6 to 24 minutes of development. Fig. 2 demonstrates that *run* stripe 1 begins to form at temporal class 1 and is formed by temporal class 2 in 100 % of embryos. On the contrary the temporal variability of formation of *ftz* stripe 4 is about 24 minutes.

Temporal variability of stripe formation is very high if to consider that the whole pair-rule pattern forms within about 50 minutes of development, which constitute the cleavage cycle 14A. Each pair-rule stripe forms with a distinctive temporal precision. Of particular interest is that the temporal precision of stripe formation is different for different stripes of one pair-rule gene.

Acknowledgements

This work is supported by NIH grant RO1 RR07801-11 and GAP award RBO-1286.

References

Akam M. The molecular basis for metameric pattern in the Drosophila embryo // Development. 1987. V. 101. P. 1–22.

Houchmandzadeh B., Wieschaus E., Leibler S. Establishment of developmental precision and proportions in the early Drosophila embryo // Nature. 2002. V. 415(6873). P. 748–9.

Jaeger J., Blagov M., Kosman D., Kozlov K., Manu, Myasnikova E., Surkova S., Vanario-Alonso C., Samsonova M., Sharp D., Reinitz J. Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster* // Genetics. 2004. (submitted).

Kosman D., Reinitz J., Sharp D. Automated assay of gene expression at cellular resolution // Pac. Symp. on Biocomput. 1997. P. 6–17.

Kozlov K., Myasnikova E., Samsonova M., Reinitz J., Kosman D. Method for spatial registration of the expression patterns of Drosophila segmentation genes using wavelets // Computational Technologies. 2000. V. 5. P. 112–119.

Myasnikova E., Samsonova A., Kozlov K., Samsonova M., Reinitz J. Registration of the expression patterns of Drosophila segmentation genes by two independent methods // Bioinformatics. 2001. V. 17(1). P. 3–12.

Reinitz J., Sharp D. Mechanism of *eve* stripe formation // Mechanisms of Development. 1995. V. 49. P. 133–158.

A LANGUAGE FOR MODELING GENETIC REGULATION IN PROCARYOTES

Tarasov D.S.[1], Leontiev A.Yu.[2], Akberova N.I.[1]*

Kazan State University, Kazan, Russia; Kazan State Academy of Veterinary Medicine
* Corresponding author: e-mail: denis@mi.ru

Keywords: *genetic regulation, computer modeling, cell device, representation format*

Summary

Motivation: Many representation forms were proposed for models of genetic regulation as well as for data storage in genetic regulation databases. However most of them lack significant features such as suitability for both modeling and database storage and ability to present various kinds of information relevant for genetic regulation studies.

Results: In this work we propose a special modeling language based on cell device architecture model. This language combines the representation power and possibility of direct usage for modeling purposes.

Introduction

Computer data processing in growing genomics databases requires a unified approach for representing knowledge on genetic processes and their regulation, along with a methods of modeling genetic processes. Various representation formats were proposed for storing sequence information (Fenyo, 1999), modeling metabolic pathways, metabolic knowledge bases (Karp *et al.*, 1996) and graphical representation of complex biological systems (Cook *et al.*, 2001). However no one of them became a standard yet and enormous complexity of living systems prevents creation of general purpose knowledge representation format, suitable for genetic databases, modeling, and sequence analysis. Furthermore, complex formal representation formats are unlikely to be used by molecular biologists. Comparisons with electrical circuit diagrams are usually ignoring the fact that they are artificial representations of artificial objects of known organization principles. In case of living cell, we have to deal with natural objects of poorly defined or unknown organization principles and all our representation formats are not native formats for living cell.

However, it was proposed that a living cell itself can be considered as computational device (Ji, 1999a, b) that uses a special language in order to store and express genetic information. This way the nature presents us a natural standard for describing living systems. Thus the study of information processing principles used by living cell can be useful for developing formats for representing biological knowledge.

In this work we propose a language for describing genetic processes in prokaryotic cells in terms of behavior of molecular biological computational device (Tarasov *et al.*, 2003).

Methods and Algorithms

Our work is based on assumption that living cell itself uses some kind of programming language in order to store and express genetic information. The problem is to determine the true structure of cell language. Assuming that some kind of cell language exist, how we are going to describe it?

As it was stated before, programs written in cell language are executed by cell device. Cell device should be considered as a useful abstraction that represents a set of molecules and their interactions that is responsible for realization of cell logic (Tarasov *et al.*, 2002). By determining principles of cell device design we can then make useful assumptions about structure and organization of cell device programming language.

Consequently, next question arises from this point: how should we study the design of cell device? As far as we are concerned, no unified procedure of any kind exists for solving such problem. Hence it would be reasonable to use some general purpose problem solving techniques.

We begin from generating possible solutions (i.e. different models of cell device) and then testing them by their ability to describe genetic regulation processes known from experimental studies. By repeating such procedure we can hope to acquire a model good enough for practical purposes.

For such approach to be successful we need some meaningful criteria to distinguish “good models” from “bad models”. In our work following criteria were used: 1. The model should allow execution and successful completion of any computational process found in living cell; 2. Any object found in model should clearly correspond to appropriate molecular object in living cell; 3. The model must not contain any object, process or principle that directly violate known principles of molecular biology; 4. The model should be useful for practical purposes.

After determining the architecture of cell device, we can design a programming language for presenting cell device programs.

For cell device programming languages to be of any practical use they must, as computer languages, use different levels of abstraction. Cell directly “understands” programs written in biopolymer sequences of nucleotides or amino-acids. Such languages are not directly understandable by humans or computers and can not be used for practical purposes. Therefore we need another special language to write cell device programs.

Implementation and Results

Enhanced principles of cell device architecture were proposed:

1. Nucleic acids, proteins and other intracellular compounds act together to form some kind of cell device. Cell device consist of data pools, passive units and active units.

2. Active objects are processes and metaprocesses. Cell device can contain a number of processes and metaprocesses.

3. Both processes and metaprocesses can be the sources of activity in the system.

4. Active units contain both data and programs written in cell language by means of their chemical structure. Passive units contain only data.

5. Processes can bind, transform and release any units. This is done according to their embedded program.

6. Metaprocesses are activated by presence or absence of particular processes in the systems. When activated they can launch additional processes according to their embedded programs.

7. All units are placed in data pools. Units from different pools can not interact directly but they can be transferred from one pool to another by special processes.

A version of high-level cell device programming language CDPL/HL was developed for representing genetic regulation in prokaryotes. The following are general principles and a very short and basic description of CDPL/HL. The description is incomplete in many ways, but gives a good assumption for the language.

Basic principles

1. The activity of cell device is specified by the programs written in cell device programming language (CDPL).

2. Complete CDPL programs contain several parts. These are parts responsible for describing initial pool state, data types, embedded programs for processes and embedded programs for metaprocesses.

3. CDPL programs are object oriented and usually can have both declarative and procedural meaning. Each statement in active units embedded programs is of the form:

<action> :- <condition>. Other statements are descriptive and concern description of either initial pool state or data types.

Structure of the program. CDPL/HL uses a prolog-like syntax and presents a high level of abstraction of genetic regulation processes. It differs from early proposed CDPL-1 like the high level computer language differs from assembler codes. Like assembler, CDPL-1 can be used from CDPL/HL programs. A few keywords and syntax elements are used. Molecular biology terminology is generally (but not always) preferred to computer programming or basic cell device terminology.

A program begins from declaring domains of gene expression (operons), followed by description of promoter structure for each operon. Statements in section "Promoters" defines conditions of promoter activation. Section proteins describes a proteins that are synthesized.

Example. The following example describes the regulatory cascade of bacteriophage lambda.

Declare

domain_1:- N,cIII,int,xis

domain_2:- Cro,cII,O,P,Q

Promoters *// Description of genetic structure*

Pl:n1 :- **not**(Cro:n2),not(cI:n3). Pr:N1 :- **not**(Cro:n2),not(cI:n3).

Pe:n4:- cII:n5,cIII:n6. Pm :- cI:n7,not(Cro:n8). lytic_genes :- Q:N9.

Pr1.

Proteins *// Description of proteins and their effects on regulation*

cI :- Pe;Pm.

case domain_1/1 :- Pr **or** domain_1 :-Pr,N:n9.

case domain_2/1 :- Pl **or** domain_2:-Pl,N:n11.

Cro :- Pr. N:- Pl.

lytic_genes :- Pr1,Q:n11.

Discussion

A model of cell device architecture gives as a basic understanding of data representation and computational "standards" found in prokaryotic cell. It may suggest a way in which experimentally obtained data on genetic regulation should be interpreted. Many languages for cell computational device programming can exist, including cell native biopolymer sequence based languages and its artificial high-level abstractions that are understandable by humans and computers but have generally the same structure with native language of prokaryotic cell.

The proposed CDPL/HL version of cell device programming language gives us a very compact representation of genetic regulation processes. The phage lambda development cascade, shown above, requires a 13 lines of CDPL/HL code and for about four printed pages when explained in plain English. Thus, CDPL/HL programs are compact and easily understandable with a little practice.

These programs can also be used for modeling complex genetic regulation, using cell device simulator – a computer program, that simulates the behavior of cell device. CDPL/HL translator is now under construction.

References

Cook D.L., Farley F.J., Tapscott J.S. A basis for a visual language for describing, archiving and analyzing functional models of complex biological systems // Genome Biol. 2001. V. 2(4). P. 1–10.

Fenyo D. The biopolymer markup language // Bioiformatics. 1999. V. 15(4). P. 339–340.

Ji S. The cell as the smallest DNA-based molecular computer // Biosystems. 1999a. V. 52. P. 123–133.

Ji S. The linguistics of DNA: words, sentences, grammar, phonetics and semantics // Ann. New York Acad. Sci. 1999b. V. 870. P. 411–417.

Karp P. Rilley M., Palley S., Pelligrini-Toole A. EcoCyc: electronic encyclopedia of *Esherichia coli* genes and metabolism // Nucleic Acids. Res. 1996. V. 24(1). P. 32–40.

Tarasov D.S, Akberova N.I., Leontiev A.Yu. Architecture of cell device // Proc. of BGRS'2002. 2002. V. 3. P. 216–218.

Tarasov D.S., Akberova N.I., Leontiev A.Yu. The model of biological molecular computational device and its application to automatic genome annotation // Proc. of MCCMB'2003. 2003. P. 225–226.

PRINCIPLES OF ORGANIZATION AND LAWS OF FUNCTIONING IN CONTROL GENE NETWORKS

Tchuraev R.N.

Institute of Biology, Ufa Research Center, Russian Academy of Sciences, Ufa, Russia,
e-mail: tchuraev@anrb.ru

Keywords: *control gene networks, cellular automata and cell ensembles, honeycomb structures, metastability*

Summary

Motivation: Knowledge of general principles of organization and laws of functioning in control gene networks makes possible the elaboration of efficient algorithms based on them to solve concrete analysis and synthesis problem**s.**

Results: The paper presents a theoretical mathematical model developed on the basis of strictly formulated premises.

Introduction

In this research objects are represented as intracellular control systems whose function is to control rapid metabolic and slow ontogenetic processes. The systems are put into correspondence with self-reproducing cellular automata and their ensembles in such a way that each individual ensemble of cellular automata corresponds to an individual eukaryote. The theory seeks to get well-proved assertions important for understanding biological aspects of the mechanisms, by which hereditary information is stored, encoded and transmitted, and the way it is realized in ontogenetic processes of self-reproducing multicellular organisms.

From Equations of Gene's Activity to Cellular Automata. For the elements of a cellular control subsystem, we consider gene blocks G_j, i.e., gene *j* taken in combination with the mechanisms of transcription, processing, transport and depot of its final product. Signal transmission from one *G*-block to another is accomplished by regulatory molecules of different specificity. Five postulates were accepted for them in the context of microapproach (Tchuraev, 1975; Tchuraev, 1998) that enabled us to derive general-form equations of activity dynamics for control (c) eukaryotic (e) gene networks $\mathrm{S}_{\mathrm{e}}^{\mathrm{c}}(G)$ represented as a finite loaded oriented graph with a set of gene blocks G = $\{\mathrm{G}_1, ..., \mathrm{G}_i, ..., \mathrm{G}_N\}$:

$$\mathbf{\Gamma}(t) = \boldsymbol{F}\ \{\boldsymbol{f}\ [\ \mathbf{\Gamma}(t-\tau), \boldsymbol{E}\ (t-\tau)]\}, \qquad (1)$$

where $\mathbf{\Gamma}(t) = \left\langle \gamma_1(t), \gamma_2(t), ..., \gamma_{\mathrm{j}}(t), ..., \gamma_N(t) \right\rangle$ is the γ-vector of the activities of all elements in the network $\mathrm{S}_{\mathrm{e}}^{\mathrm{c}}(G)$, in this case $\gamma = \gamma(\mathrm{t})$ is the binary value; *t* is the discrete time; **F** is the column dimension $N \times 1$, whose elements are Boolean functions ("composition" of logic structures); $\boldsymbol{f} = \| f_{\mathrm{ij}} \|$ is the matrix dimension $\mathrm{J}^{\mathrm{j}} \times N$, where each element f_{ij} is a restrictedly determined operator that connects the internal variable υ_{ij} to a sequence of input signals e_{ij} entering via the *i*-th input channel of the *j*-th gene block; $\tau = \max \tau_{ij}$ is the maximum delay of output signals among all *G*-blocks affecting the given one (G_i);

$\boldsymbol{E}(t-\tau) = \left\langle e_1(t-\tau_1)\, e_2(t-\tau_2) ... e_h(t-\tau_h) ... e_H(t-\tau_H) \right\rangle$ – 0,1 word of length *H*, in this case *H* is the number of input channels in the network $\mathrm{S}_{\mathrm{e}}^{\mathrm{c}}(G)$.

The magnitude of the γ-vector components can be experimentally observed by noting presence or absence of specific gene products at a given instant of time. Hence $\mathbf{\Gamma}$(t) is *the observable behaviour of the control gene network*. Following from the Kobrinsky-Trakhtenbrot theorem (1962), it may be asserted (1) that *if stationary sequences of input signals enter via all inputs of any finite gene subnetwork, the latter is either at rest or in the periodic regime*. The number of points of rest and periodic regimes are finite values.

It is known that the restrictedly determined operator with finite weight realized by the network $S_e^c(G)$ can be presented as a finite automaton; we call it the *cellular automaton* $A_e^c(G)$, in which the internal structure is formed by the control cellular network $S_e^c(G)$.

Principles of Organization in Control Gene Networks

Since each specific regulatory substance in the cell is a product of gene blocks, we accept the following *structural postulate*:

For each element G_j of the network $S_e^c(G)$ there is such an element G_i that at least one of its outputs is connected to the input of the element G_j by means of communication channel.

Based on the structural postulate and finiteness of elements in the network $S_e^c(G)$, assertion (2) can be stated in the following way.

Each gene block (element) of control gene network $S_e^c(G)$ adheres to at least one control loop (oriented cycle), i.e., it either enters the oriented cycle or is connected to this cycle with a signal circuit.

In any event eukaryotic cellular control gene networks are arranged in a *modular fashion*, i.e., any eukaryotic gene network S_e^c can be represented as a network of blocks (subnetworks), where each block at a higher level of complexity is constructed with the blocks from the previous level of complexity, both levels functioning as an entity.

As follows from this principle, the control gene networks $S_e^c(G)$ *can be transformed into the network* $\tilde{S}_e^c(G)$ *whose elements will consist of the following modules: genetic triggers (bistable memory modules), oscillators and delay logical combinators.*

There is time hierarchy in the structures of cellular control gene networks $S_e^c(G)$, which ensures gene sequential switching during ontogenesis. As follows from assertion (2), there should be feedbacks in gene networks $S_e^c(G)$. Does the existence of these feedbacks exclude a hierarchic principle in the organization of cellular control gene networks $S_e^c(G)$? No, it apparently does not, if, as applied to networks $S_e^c(G)$, the term *heterarchy* will mean the existence of feedbacks that connect output channels of genetic blocks at different hierarchic levels to input channels of gene blocks of a higher hierarchic rank. Summing up, the following assertion can be stated: *any cellular control gene networks are arranged on principles of both hierarchy and heterarchy.* In addition, a correspondence is found between the structure of control gene network $S_e^c(G)$ and schematic blocks of the hereditary program that realizes the inherited algorithm χ during ontogenesis (*principle of correlation between structure and ontogenetic function*).

It is known that complicated multimeric complexes, agregulons, whose specificity depends on composition, may serve as carriers of molecular signals (Jacob, 1993). Using n number of different monomers, q number of different regulatory multimers can be formed:

$$q = \frac{n!}{(n-p)!\,p!}, \tag{2}$$

where p is the level of multimers. Thus, *there are combinatorial modules in eukaryotic control networks, and their function consists in generating a large number of different signals from a small number of molecular signals.*

Laws of Functioning in Cellular Automata

A canonical description of the cellular automaton $A_{\mathrm{e}}^{\mathrm{c}}$ has the form:

In the general form the cellular automaton $A_{\mathrm{e}}^{\mathrm{c}}$ is described with five symbols (E,∇, Ω, Φ, Ψ), where E and ∇ are the input and output alphabets, Ω is the set of internal memory states Ξ, Φ and Ψ denote the transition and output functions, respectively.

Hence we get the following description of the discrete finite automaton $A_{\mathrm{e}}^{\mathrm{c}}$. *The input alphabet* E of the automaton $A_{\mathrm{e}}^{\mathrm{c}}$ with n_1 number of input channels constitutes a set of corteges (words of length n_1): E = {**e**}, where $\mathbf{e} = \langle \varepsilon_1(\mathrm{t}), \varepsilon_2(\mathrm{t}), \ldots\ \varepsilon_l(\mathrm{t}), \ldots\ \varepsilon_{\mathrm{n}1}(\mathrm{t})\rangle$, $l = \overline{1, n_1}$, and the elements ε of the cortege **e** are the binary values. *The output alphabet* ∇ of the automaton is $A_{\mathrm{e}}^{\mathrm{c}}$ best represented as a set of γ-vectors of the activities $\mathbf{\Gamma}(t) = \left\langle \gamma_1(t), \gamma_2(t), \ldots, \gamma_{\mathrm{j}}(t), \ldots, \gamma_N(t) \right\rangle$, where $\gamma_j = \gamma_j(\mathrm{t})$ denoting the activity of the gene block G_{j} are the elements of control gene network $\mathrm{S}_{\mathrm{e}}^{\mathrm{c}}(G)$. In other words, at each discrete instant of time *t* it is possible to record the *observable* values, or the activities of all genes in the control gene network $\mathrm{S}_{\mathrm{e}}^{\mathrm{c}}(G)$, i.e., for the output channels of the automaton $A_{\mathrm{e}}^{\mathrm{c}}$ we consider the output channels of all its elements, not only those unconnected to the other elements of the network. Such a representation of output symbols in the cellular automaton is motivated by a possibility to have experimentally observable patterns of gene activities in the control gene network judging, for example, by the presence (or the absence) of primary transcripts. Thus, the output alphabet ∇={Γ} of the cellular automaton $A_{\mathrm{e}}^{\mathrm{c}}$ represents a set of all possible words **Γ** of length *N*:

$\nabla = \{\Gamma_1, \Gamma_2, \ldots, \Gamma_j, \ldots, \Gamma_{2^N}\}$, where $\mathbf{\Gamma} = \mathbf{\Gamma}(t) = \left\langle \gamma_1(t), \gamma_2(t), \ldots, \gamma_{\mathrm{j}}(t), \ldots, \gamma_N(t) \right\rangle$ is the γ-vector of gene activities in the control gene network. *A set of states* **Ω** in the memory Ξ of the automaton $A_{\mathrm{e}}^{\mathrm{c}} : \Omega = \{\omega_0, \omega_1, \ldots, \omega_m, \ldots, \omega_{\mathrm{M}}\}$, $m = (\overline{1, \mathrm{M}})$.

It should be noted that if one takes into account the internal structure of the automaton $A_{\mathrm{e}}^{\mathrm{c}}$, each of the states ω∈Ω will constitute a complex composition of ministates in each element of the network $\mathrm{S}_{\mathrm{e}}^{\mathrm{c}}$, the set of states Ω being the Descartian product of ordered sets of these states N.

$$\Omega = \mathop{D_{\otimes}}_{i=1}^{N} Q_i, \tag{3}$$

where symbol $D_{\otimes}$ denotes the Descartian product, and Q_i is the set of states in the i-th element of the network S_e^c, in this case $i = \overline{1, N}$.

The transition Φ *and output* Ψ *functions* of the automaton A_e^c have the form:

$$\omega(t+1) = \Phi\,[\omega(t), \mathbf{e}(t+1)],$$

$$\Gamma\,(t+1) = \Psi[\omega(t+1)], \tag{4}$$

The cellular automaton A_e^c described in such a way is the model of the control gene network in the most general form.

Let us introduce the act of reduplicating: $A_e^c \to {}'A_e^c \cup {}''A_e^c$ for cellular automata A_e^c , where ${}'A_e^c$, ${}''A_e^c$ are the copies of the parent automaton A_e^c (Tchuraev, 1991). The networks S_e^c, ${}'S_e^c$ and ${}''S_e^c$, and "parent" and "daughter" automata are isomorphic.

During the reduplication of the automata A_e^c there occurs a sequential formation of the *cellular automata ensemble* A_e^x, which corresponds to individual x that converts itself from a zygote into a duplicated form. A natural requirement for the ensemble A_e^x is its ability for self-duplication. Furthermore, generalized cellular control gene networks (CGN) in eukaryotes have three fundamental properties: a) during cell sequential divisions there must be generative cells, in which the CGN return in their initial state; b) after a series of initial cell (zygote) divisions the CGN should be able to give rise to several "somatic" cell lines (*ability for divergent determination and differentiation*); c) some functional states of the CGN should be preserved in a series of cell sequential divisions (*stability of determinate states*). Within the model of the cell ensemble A_e^x these properties are accepted as the premises.

For two-dimensional cellular automata, Edward F. Moore (1962) introduced a honeycomb neighbourhood ("universe") defined by six positions, one of which is the causality principle. In Moore's formal language the cell ensemble A_e^x is put into correspondence with the honeycomb-like block x and its properties (a-c) are formalized.

As a result of the formalized *hypothesis for the ability of developing cell ensembles to be self-differentiated* in the neutral medium, we have established a theorem proving the existence of specific *metastable states* $\breve{\omega}$ within a multitude Ω of states in the cellular automata A_e^c , such that:

$$A_e^c / \breve{\omega} \to {}'A_e^c \;/\; \omega_\upsilon \cup {}''A_e^c / \;\omega_p \text{, where } \omega_\upsilon \text{ and } \omega_p \text{ are the different states.} \tag{5}$$

The existence of metastable states imposes constraints on the structures of control gene networks. By way of example, we have studied the behaviour of the simplest networks in the metastable state.

Let us write down the behaviour laws of cellular automata. According to assertion (1), the 1st law of the activity dynamics has the following form, in the context of microapproach, for the observable behaviour of the cellular automaton:

$$\mathbf{\Gamma}(t) = \Psi\{\Phi\ [\boldsymbol{\omega}(t\text{-}1),\ \mathbf{e}(t)]\} = \tilde{\mathbf{\Gamma}}\ ,\ \text{if}\ \boldsymbol{\omega}(t - 1) = \tilde{\boldsymbol{\omega}}\ ,\ \mathbf{e}(t) = \tilde{\mathbf{e}}, \tag{6}$$

where $\mathbf{\Gamma}(t) = \left\langle \gamma_1(t), \gamma_2(t),..., \gamma_j(t),..., \gamma_N(t) \right\rangle$ is the γ-vector of the activities of gene blocks in the control gene network (subnetwork) $S_e^c(G)$, Ψ is the single-valued function of state $\boldsymbol{\omega}$, Φ is the transition function being not necessarily single-valued at some points, $\tilde{\boldsymbol{\omega}}$ - stationary state, $\tilde{\mathbf{e}}$ - neutral input symbol.

The 2nd law of the activity dynamics:

If $\mathbf{e}(t)\ \tilde{\mathbf{e}} =$and $\boldsymbol{\omega} \neq \tilde{\boldsymbol{\omega}}$, the function Φ (and accordingly Ψ) are periodic.

Thus, in the neutral medium any cellular automaton and associated intracellular control network are either in one of the states of rest or in one of possible periodic regimes.

The 3rd law of the activity dynamics:

If $\mathbf{e}(t) \neq \tilde{\mathbf{e}}$ and $\boldsymbol{\omega} \neq \tilde{\boldsymbol{\omega}}$,

$$\mathbf{\Gamma}(t) = \Psi\{\Phi[\boldsymbol{\omega}_{i_\alpha}(t\text{-}1),\ \mathbf{e}(t)]\} = \begin{cases} \mathbf{\Gamma}_\alpha \\ \mathbf{\Gamma},\ \mathit{if}\ \ \mathbf{e} \neq \mathbf{e}_\alpha \end{cases}, \tag{7}$$

where $\boldsymbol{\omega}_{i_\alpha}$ is the state competent to signal $\mathbf{e}_\alpha$. This establishes the relation of mutual specificity between states and signals.

As a result of the existence of metastable states (5), we derive the 4th law of the activity dynamics:

$$\Gamma(t) = \Psi\{\Phi[\breve{\boldsymbol{\omega}}(t - 1), \tilde{e}\ (t)]\} = \begin{cases} \mathbf{\Gamma}_\upsilon, \\ \mathbf{\Gamma}_\rho, \end{cases} \tag{8}$$

where designations have the same meaning as in expressions (5) and (6).

Expressions (5) and (8) imply the existence of such cell division during ontogenesis, when, in the neutral extracellular medium, daughter cells will differ in the activity of at least one gene (divergent determination). It can be hoped that this theory suggests an answer to the long-time question: whence the order appears during ontogenesis?

References

Jacob F. Du répresseur à I'agrégulat // C.R.Acad. Sci. Paris: Sciences de la vie/Life sciences, 1993. V. 316. P. 331–333.

Kobrinsky N.E., Trahtenbrot B.A. Introduction into the finite automata theory // Phys. Math. Giz. M., 1962. (In Russian).

Moor E.F. Mathematical models of self-reproduction // Proc. of Symposia in applied mathematics. XIV. American Mathematical Society Press, Providence, 1962. P. 36–62.

Tchuraev R.N. Mathematic-logical models for molecular control systems / Ed. V.A. Ratner. Studies on Mathematical Genetics, ICG Press, Novosibirsk, 1975. P. 67–76. (In Russian).

Tchuraev R.N. 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. 1991. V. 151. P. 71–87.

Tchuraev R.N. The equations of dynamics of genes activities in a general view // Proc. of BGRS'1998. Novosibirsk, 1998. P. 128–131.

MODELING AND CONSTRUCTION OF MOLECULAR TRIGGER IN *E. COLI*

***Vasilenko N.L.*[1], *Balueva K.E.*[2,4], *Likhoshvai V.A.*[1,3], *Nevinsky G.A.*[2], *Matushkin Yu.G.**[1]**

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Institute of Chemical Biology and Fundamental Medicine of SB RAS, Novosibirsk, Russia; [3] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia; [4] Novosibirsk State University, Novosibirsk, Russia
* Corresponding author: e-mail: mat@bionet.nsc.ru

Keywords: *hypothetical gene network, mathematical model, computer model, regulation, stationary solutions, stability, critical points, theory of gene networks, trigger, inducer, repressor*

Summary

Motivation: Design of genetically engineered constructs that would allow for controlling operation of certain genes is among the most topical problems in the modern biology. We are developing the theory of modeling hypothetical and actual gene networks, which provides prediction of the main modes of gene network operation. The goal of this work was to integrate theoretical and experimental approaches to analysis of the operation patterns of relatively simple vector system with trigger properties.

Results: Behavior of simplest gene networks with trigger and cyclic operation modes was studied qualitatively and quantitatively using mathematical models (portrait and regulatory circuits). Functioning of trigger plasmids under the effect of inducers in *E. coli* cells was verified experimentally. Optimal conditions for studying the dependence of fluorescent signal intensity on the presence/absence of inducer and duration of its action on repressor were selected.

Introduction

Earlier, we developed a method for modeling GNs (Likhoshvai *et al.*, 2000; Likhoshvai *et al.*, 2001a) and several computer models (Bazhan *et al.,* 1995; Ratushny *et al.,* 2003; etc.) as well as the fundamentals of the theory of hypothetical gene networks (HGNs), which are actually the models of regulatory circuits of real GNs (Likhoshvai *et al.*, 2001b, 2003, 2004; Likhoshvai, Fadeev, 2003). Mathematical tools for studying HGNs were proposed. It was demonstrated that HGNs could have stable stationary and cyclic operation modes as well as attractors of a more intricate nature. Numerical calculations showed that only the structure–function relations determine the limit properties of canonical HGNs at sufficiently large values of synthesis and repression parameters. A number of empirical criteria connecting the limit properties of HGNs with the properties of the corresponding structural graphs were formulated. With reference to the real gene networks, this means that a certain minimal complexity (nonlinearity) of the processes regulating the activities of GN elements is the necessary condition for existence of the necessary number of stationary and/or cyclic GN operation modes. The necessary complexity may be reached by multimerization of repressor proteins and/or occurrence of a sufficiently large number of intermediate stages between the gene and the regulatory protein.

Research into HGNs opens the possibility for synthesizing gene networks with any prespecified number of stationary and/or oscillating modes as well as for realizing many other, possibly, more intricate systems with operation modes oriented to solving practical problems.

Results and Discussion

Models. We constructed and numerically studied models of self-regulating operon (SRO), molecular trigger (MT), and molecular oscillator (MO) composed of three operons successively repressing one another. The models were constructed using generalized chemical method of modeling (Likhoshvai *et al.*, 2000). It was demonstrated that a region of at least 2000 bp between promoter and protein coding part is necessary for the onset of self-oscillations in SRO. MT and MO function according to experimental data (Elowitz, Leibier, 2000; Fig. 1).

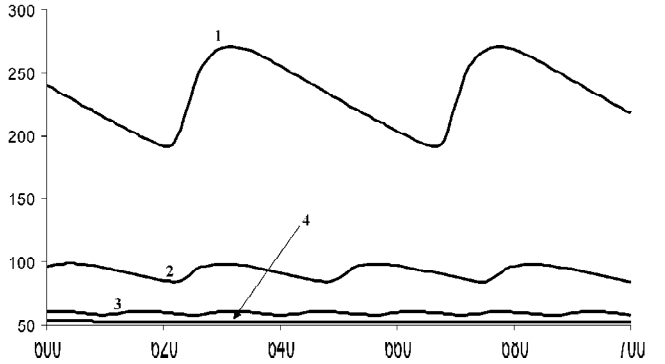


Fig. 1. Self regulation of operon expression efficiency: 1, insertion with a length of 10 000 nucleotides (the period T = 46 min); 2, insertion with a length of 5000 nucleotides (the period T = 26 min); 3, insertion with a length of 3000 nucleotides (the period T = 15 min); 4, insertion with a length of 2000 nucleotides (stationary); the ordinate, concentration of the reporter protein; the abscissa, time in min.

For SRO and MO, properties of periodicity were studied. The minimal period for SRO at physiological parameters amounted to ~15 min; for MO, at least 100 min. Numerical experiments with MO demonstrated that the length of the period depended essentially on the gene copy number (Fig. 2). This suggests fluctuations in copy numbers of plasmids underlies a large variation of the periods observed (Elowitz, Leibier, 2000). In the model of MT, the mean time of changeover amounted to

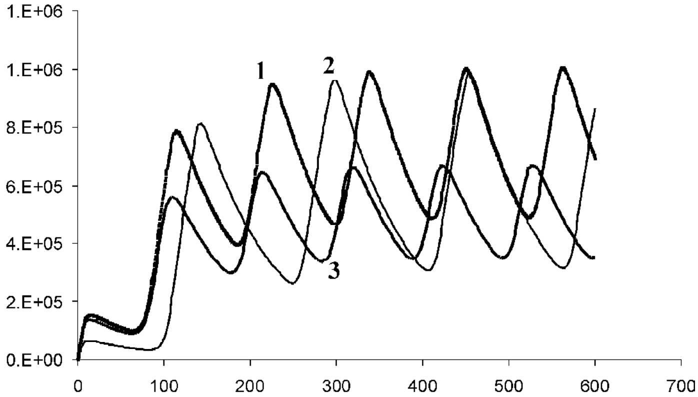


Fig. 2. A three operon oscillator: 1, the plasmid copy number is 1/100 plasmids/cell (the period T = 110 min; 2, the copy number is 10/100 plasmids/cell (the period T = 156 min; and 3, the copy number is 1/10 plasmids/cell (the period T = 104 min).

approximately 25–30 min. To verify the adequacy of the conclusions obtained using the model, we commenced experimental construction of the trigger system.

Experiment. Earlier, trigger plasmids containing two repressors and two constitutive promoters were constructed (Elowitz, Leibier, 2000; Gardner *et al.*, 2000; Tropynina *et al.*, 2002). The plasmids pTAK and pIKE were kindly provided by J.J. Collins (Fig. 1). Both plasmids contain lacI gene. Lac repressor (R2) binds to promoter Ptrc-2 (P2) to form the first pair promoter–repressor. As the second pair (P1 and R1), the promoter P_Ls1con and thermosensitive repressor cIts of λ phage are used in plasmids pTAK; tetR gene, encoding Tet repressor, binding to the promoter $P_L$tetO-1, in plasmids pIKE. Changeover in plasmids pTAK occurs when IPTG, which interacts with LacI repressor, is added or when temperature is increased. Changeover in plasmids pIKE occurs when IPTG or tetracycline (Tc) is introduced. The state of triggers was monitored according to change in the expression level of GFP protein, which was under the control of Ptrc-2 promoter.

Individual colonies of JM2.300 cells, carrying pTAK or pIKE plasmids were transferred into selective medium supplemented (if necessary) with 2 mM IPTG to grow a night culture at 37 °C (pIKE) or 32 °C (pTAK). Then, the cells were grown in either the presence or absence of the corresponding inducer. Cells were transferred into minimal medium to record the intensity of GFP fluorescence at an excitation wavelength of 480 nm in a SFM 25 (KONTRON INSTRUMENTS, Italy) fluorimeter. All the measurements were made during the logarithmic growth phase. If necessary, the medium was supplemented with tetracycline as the second inducer (pIKE) or cells were incubated at 42 °C (pTAK).

Upon a 3–6-h induction, cells were transferred into the fresh medium free of inducer. The cells retained the state when GFP was expressed in the absence of the inducer. Upon the effect of the second inducer, the triggers switched to the functional state when GFP expression was absent.

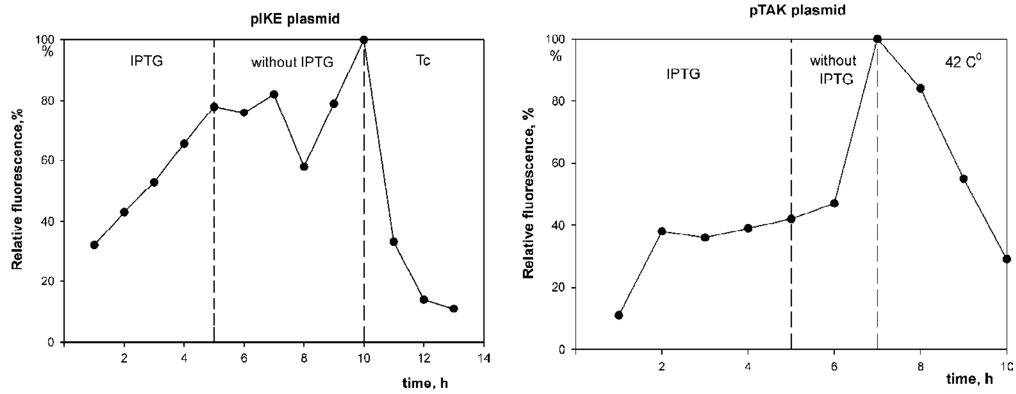


Fig. 3. Dependence of GFP fluorescence on time and presence of inducer.

This state was stably inherited by the next cell generations. Fig. 3 shows the dependence of GFP expression level on time and presence/absence of inducers.

As control experiments, cell cultures were divided into two parts. The first was grown in the presence of IPTG; the second, in the absence. Upon 6 h of growth in the presence of inducer, the cells were transferred into an inducer-free fresh medium. Both cultures retained the corresponding states (with a high or low GFP expression level) for 1 day. The ability of trigger plasmids to preserve two stable functional states in a succession of cell divisions is shown in Fig. 4.

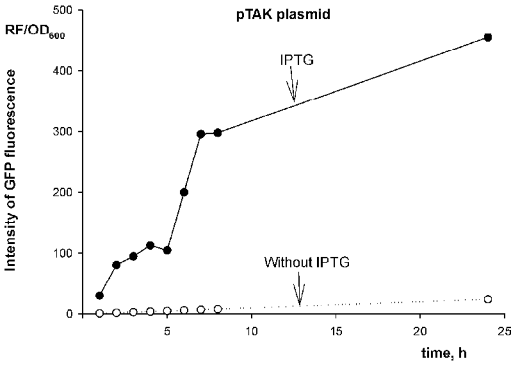


Fig. 4. Stability of functional states of the plasmid pTAK.

Thus, it is demonstrated that the trigger plasmids under certain conditions are able to preserve stably one of the two alternative functional states in the succession of *E. coli* cell divisions. Further, we plan to construct a system with cyclic operation mode as well as more complex systems combining both stationary points and cycles.

Acknowledgements

Work was supported in part by the Russian Foundation for Basic Research (No. 02-04-48802, 03-04-48829, 04-01-00458), Russian Ministry of Industry, Sciences and Technologies (Nos. 43.106.11.0011, 43.073.1.1.1501), Integration project of SB RAS No. 148, RAS Presidium Programs “Molecular and Cellular Biology” (No. 10.4) and “Origin and evolution of biosphere” (contract No. 10002-251/П-25/155-270/200404-082). Authors are grateful to J.J. Collins (Boston University, USA) for the plasmids and to T.P. Kamynina for helpful discussion.

References

Bazhan S.I., Likhoshvay V.A., Belova O.E. Theoretical analysis of the regulation of interferon expression during priming and blocking // J. of Theoretical Biol. 1995. V. 175. P. 149–160.

Elowitz M.B., Leibier S. A synthetic oscillatory network of transcriptional regulation // Nature. 2000. V. 403. P. 335–338.

Gardner T.S., Cantor C.R., Collins J.J. Construction of a genetic toggle switch in *Escherichia coli* // Nature. 2000. V. 403, P. 339–342.

Likhoshvai V.A., Matushkin Yu.G., Vatolin Yu.N., Bazhan S.I. A generalized chemical kinetic method for simulating complex biological systems. A computer model of λ phage ontogenesis // Computational Technologies. 2000. V. 5, N 2. P. 87–99.

Likhoshvai V.A., Matushkin Yu.G., Ratushny A.V., Anan’ko E.A., Ignat’eva E.V., Podkolodnaya O.A. Generalized chemical kinetic method for modeling gene networks // Mol. Biol. (Mosk.). 2001a. V. 35. P. 1972–1980.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. On the relation of gene network graph to qualitative mode of its operation // Mol. Biol. (Mosk.). 2001b. V. 35. P. 1080–1087.

Likhoshvai V.A., Fadeev S.I. On hypothetical gene networks // Sib. Zh. Industr. Matemat. 2003. V. 6. P. 134–153.

Likhoshvai V.A., Matushkin Yu.G., Fadeev S.I. Problems in the theory of operation of gene networks // Sib. Zh. Industr. Matemat. 2003. V. 6. P. 64–80.

Likhoshvai V.A., Fadeev S.I., Demidenko G.V., Matushkin Yu.G. Modeling of multistage synthesis of a substance without branching using an equation with retarded argument // Sib. Zh. Industr. Matemat. 2004. V. 7. P. 73–94.

Ratushny A.V., Likhoshvai V.A., Ignat’eva E.V., Matushkin Yu.G., Goryanin I.I., Kolchanov N.A. Computer model of the gene network of cholesterol biosynthesis regulation in the cell: analysis of the effect of mutations // Dokl. Akad. Nauk. 2003. V. 389. P. 90–93.

Tropynina T.S., Golubev O.V., Stupak E.E., Churaev R.N. Construction of artificial digenic network possessing epigenic properties // Mol. Biol. (Mosk.). 2002. V. 36. P. 605–609.

STOCHASTIC MODEL OF TRANSLATION ELONGATION BASED ON CONTINUOUS TIME MONTE CARLO METHOD

Vladimirov N.V.[*1], *Likhoshvai V.A.*[1,2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Ugra Research Institute of Information Technologies, Khanty-Mansyisk, Russia
* Corresponding author: e-mail: nikita@bionet.nsc.ru

Keywords: *translation, ribosome, codon, secondary structure, elongation, stochastic modeling, model of translation, Monte Carlo method*

Summary

Motivation: Protein synthesis is one of the most energy consuming processes in cell. As a result, there are strong evidences of optimization at different stages of mRNA translation in many organisms. In particular, for individual mRNAs their codon composition and local secondary structures are biased, resulting in high optimization of elongation. To reveal fine patterns of ribosome traffic along mRNA, development of mathematical model of mRNA translation is required.

Results: A stochastic model of elongation based on continuous time Monte Carlo method has been developed. The model simulates initiation of translation, ribosome movement along mRNA, tacking into account the current codon in A-site of ribosome, rate of cognate tRNA binding to the codon, local secondary structures and ribosome size. A new effect of periodical localization of non-optimal codons in mRNA has been predicted using the model. This effect has been shown to exist in natural *E. coli* genes. The length of the period is equal to the ribosome size obtained in experimental studies, confirming the model validity and its applicability for solving biological problems.

Introduction

The rate of protein synthesis is one of the most crucial characteristics of cell ontogenesis. The translation machinery can consume up to 50÷70 % of total cell energy (for example, *E. coli* in exponential growth phase). Therefore, enhancement in translation efficiency is a significant factor of evolution. In a number of works the usage of synonymous codons with different rates of translation was shown to be a widespread mechanism of translation optimization (Ikemura, 1985; Li, Lou, 1996). The optimal codons with high translation rates are preferred in highly expressed genes. Such codons correspond to the most abundant cognate tRNAs. It has been shown that for a number of organisms (*E. coli*, *S. cerevisiae* and others) the expression level of genes may be estimated using the frequency of optimal codons usage in their mRNAs (Li, Lou, 1996). Local complementary mRNA structures which slow down the ribosome movement may also influence the rate of translation (Likhoshvai, Matushkin, 2002).

In this work we have developed a stochastic model of translation which takes into account the rate of initiation, successive movement of ribosome along the mRNA, translation rate of current codon in its A-site, ribosome pausing at local mRNA hairpins, and steric size of ribosome. The model has been implemented in a program on the base of continuous time Monte Carlo method. The program estimates the basic characteristics of translation for given mRNAs: rate of protein synthesis, level of optimization of codon composition, average number of ribosomes moving along mRNA, fraction of ribosomes standing in queue, average distance between ribosomes in polysome. It also allows to obtain elongation speed profiles and variances of the estimates.

Recently it has been shown that modeling of translation within deterministic approach (solving systems of differential equations) can be carried out only in very crude approximation (Likhoshvai, Matushkin, 2000). We assume that stochastic methods nowadays are the only approach for detailed modeling of translation.

The model has been used for studying the patterns of elongation profiles of *E. coli* mRNAs. We have confirmed the well-known effect of frequent usage of rare codons in 5'-proximal coding regions of mRNAs (Bulmer,1991). Besides, we have predicted one previously unknown effect of periodical localization of non-optimal codons due to the queuing of ribosomes. This effect was shown to exist in real *E. coli* genes. The results are consistent with experimental studies about ribosome steric size (Kozak, 1983).

Methods and Algorithms

Coding sequences of *Escherichia coli* K12 genes have been extracted from GenBank database.

The rate constants k_{bind}^{j} of cognate tRNA binding to the *j*-th codon (*j*=1,..,61) exposed in A-site of ribosome, which determine the rate of codon translation, were calculated using the algorithm described in (Likhoshvai, Matushkin, 2002). The algorithm for k_{bind}^{j} calculation uses relative frequencies of codon usage in a sample of genes, which is automatically composed by certain rules.

Elementary events in the model are the following: a) binding of the *i*-th ribosome to the start codon; b) binding of cognate aa-tRNA to the codon exposed in A-site of the *i*-th ribosome; c) translocation of ribosome to the next codon (if next ribosome does not block the way); d) termination of translation.

In the implementation of the model the continuous time Monte Carlo method with the algorithm described in (Gibson, 2000) has been used.

The influence of local mRNA hairpins is considered through increasing the time point of ribosome translocation t_{trans}^{i} on small value Δt_{LCI}, which has a physical meaning of ribosome pausing for untwining the hairpin. The value Δt_{LCI} is a random variable with probability density:

$$p_{LCI} = \int_{0}^{LCI} \frac{k^{n+1}x^{n}}{G(n+1)} e^{-kx} dx,\ k = m/\sigma^2,\ n = (m/\sigma)^2.$$

Here m and σ are the mean and variance of gamma-distribution, respectively, and LCI (local complementarity index) is the measure of self-complementarity of mRNA region before the ribosome (Likhoshvai, Matushkin, 2002).

Implementation and Results

The program is implemented in C++. Ribosomes are independent objects with a set of states (current position on mRNA, type of codon in A-site, etc.) and pointers to the neighboring ribosomes. Polysome is implemented as dynamic list of ribosomes.

Calculation of elongation profiles for a sample of *E. coli* K12 genes confirms the previously known effect of frequent usage of non-optimal codons in the beginning of *E. coli* genes (Bulmer 1991) (Fig. 1). This result supports the validity of calculation algorithm for translation rate constants of individual codons (Likhoshvai, Matushkin, 2002).

We have carried out a numerical analysis to investigate how one non-optimal codon influence the evolutional drift of other codons in its 5'-proximal region of coding sequence. We speculated as follows. Let us consider an mRNA consisting of optimal codons and one non-optimal among them. A ribosome will pause on this non-optimal codon, waiting for cognate tRNA. It is obvious that the following ribosomes will form a queue. The A-sites of queuing ribosomes will be separated from each other by a distance approximately equal to the size of ribosome. The codons in these A-sites will be exposed to weaker natural selection pressure by optimality because changing them to optimal ones will not increase the overall rate of mRNA translation (ribosomes pause on them anyway). Therefore, we can suppose that in the 5'-proximal region of coding sequence from the non-optimal codon there will appear new non-optimal codons due to mutation process. In this case the distance between these non-optimal codons will be approximately equal to the size of ribosome. The results of simulation for hypothetic mRNA are shown in Fig. 2.

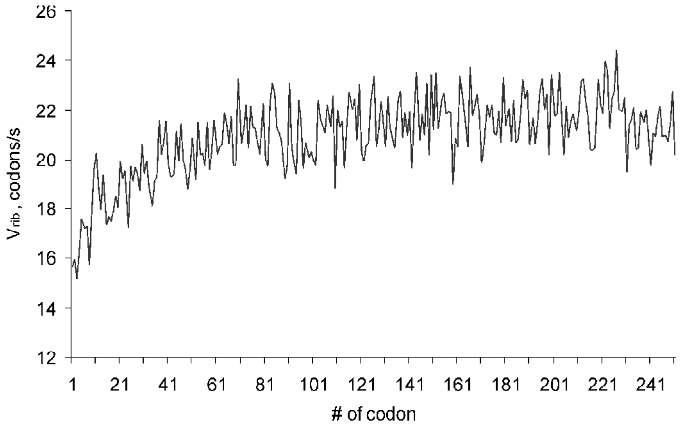


Fig. 1. Averaged elongation profile for mRNAs of highly expressed *E. coli* genes. The horizontal axis shows the number of codon in the ORF (#1 is the codon next to the start one). The vertical axis (V_{rib}) is the average speed of ribosome (codons/sec).

To verify our reasons we have analyzed average translation rates of codons in highly expressed genes of *E. coli* K12, and indeed revealed this effect (Fig. 3). For this purpose we have taken 90 genes with the highest frequency of optimal codons usage, and considered the fragments consisting of the first 50 codons (from #1 to #50). In these fragments we have selected non-optimal codons and calculated average translation rates of all codons at distance of D=2,...d_{max} to the left from the selected ones, in their 5'-proximal regions (d_{max} is the distance between selected non-optimal codon and the start one). Results are shown in Fig. 3.

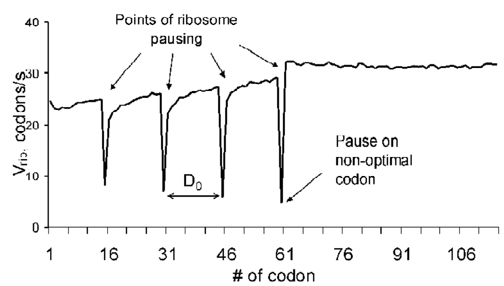


Fig. 2. Elongation profile of hypothetic mRNA consisting of optimal codons with high translation rate, and one non-optimal in position 60. D_0 is the putative size of ribosome. Axes are the same as in Fig. 1.

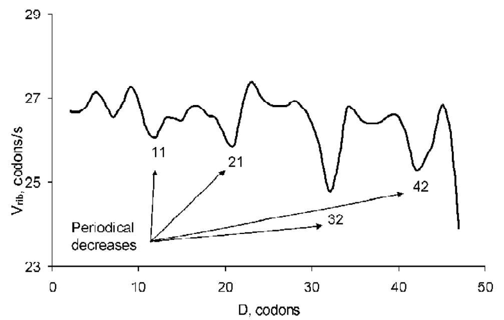


Fig 3. Average translation rates of codons in the 5'-proximal regions of non-optimal codons. Horizontal axis D is the distance from a non-optimal codon, vertical axis is the average translation rate of codons.

The length of period D_0 obtained is 11±1 codons. The decreases of average translation rates of codons repeat at the distances of 11, 21, 32, 42 codons (D_0, $2D_0$, $3D_0$, $4D_0$, respectively) to the left from non-optimal codons (Fig 3), in accordance with the effect predicted on hypothetic mRNA (Fig. 2). The length of period (33±3 nucleotides) with accuracy to 3 nucleotides is equal to the steric size of ribosome from experimental studies (Kozak, 1983) – about 30 nucleotides. This result sustains the validity of our hypotheses and the applicability of the model for solving biological problems related to codon biases.

The model developed may be applied for solving problems of evolution and bioinformatics related to optimization of codon composition of genes.

Acknowledgements

The work was supported by the grants No. 03-07-96833, 03-04-48829, 03-04-48506, 03-01-00328, 02-07-90359, 02-04-48802 of the RFBR, by the Interdisciplinary grant No. 119 and the Project No. 10.4 of the RAS Presidium Program “Molecular and Cellular Biology”, Integration project of SB RAS No. 148.

References

Bulmer M. The selection-mutation-drift theory of synonymous codon usage // Genetics. 1991. V. 129. P. 897–907.

Gibson M.A. Computational methods for stochastic biological systems. PhD Thesis. California Inst. Technology. 2000.

Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms // Mol. Biol. Evol. 1985. V. 2. P. 13–24.

Kozak M. Comparison of initiation of protein synthesis in prokaryotes, eukaryotes, and organelles // Microbiol. Rev. 1983. V. 47. P. 1–43.

Li and Luo. The relation between codon usage, base correlation and gene expression level in *Escherichia coli* and Yeast // J. Theor Biol. 1996. V. 181. P. 111–124.

Likhoshvai V.A., Matushkin Yu.G. Computer model for analysis of evolutionary drift of synonymous codons along mRNA // Computational Technologies. 2000. V. 5. Special Issue. P. 57–63.

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

RESTRICTION SITE TAGGED PASSPORTS AND MICROARRAYS FOR ANALYSIS OF COMPLEX BIOLOGICAL SYSTEMS

Zabarovsky E.R.[*1,2], Kashuba V.I.[1], Li J.[1], Kutsenko A.S.[1], Protopopov A.I.[1], Petrenko L.[1], Wang F.[1], Senchenko V.N.[3], Kadyrova E.[4], Zabarovska V.I.[1], Muravenko O.V.[2], Zelenin A.V.[2], Kisselev L.L.[2], Winberg G.[1], Ernberg I.[1], Braga E.[4], Lerman M.I.[5], Klein G.[1]*

[1] MTC and CGB, Karolinska Institute, Stockholm, Sweden; [2] EIMB, RAS, Moscow, Russia;
[3] Bioengineering Center, RAS, Moscow, Russia; [4] Russian State Genetics Center, Moscow, Russia;
[5] FCI-NCI, Frederick, USA
* Corresponding author: e-mail: eugzab@ki.se

Keywords: *genome scanning, CpG islands, NotI sites, epigenetics, tumor suppressor genes, methylation, genomic microarrays, biodiversity, NotI-tags, human gut*

Summary

Motivation: The main aim is to develop methods (experimental and computational) for the fast analysis of prokaryotic and eukaryotic genomes and complex biological systems like human gut microbial flora and cancer cells.

Results: We have developed novel tools for genome analysis (patent pending): restriction site tagged (RST) microarrays and restriction site tagged sequences (RSTS or passporting). Using NotI enzyme we have shown that NotI microarrays offer a powerful tool with which to study carcinogenesis. Moreover, NotI microarrays are the only existing microarrays giving the opportunity to detect simultaneously and differentially copy number and methylation changes. Thus they allow to check cancer cells for genetic and epigenetic abnormalities. For microbial identification NotI microarrays and passporting are significantly more specific and powerful than sequencing of 16S ribosomal genes or any gene specific microarrays. They allowed identification in bacterial mixtures thousands of known and novel microbial species and strains. Pilot experiment showed that human intestine contains more than 90 % of unknown bacteria.

Introduction

The presence of genetic alterations in tumors is widely accepted, and explains the irreversible nature of tumors. However, now, DNA methylation in CpG sites is known to be precisely regulated in tissue differentiation, and is supposed to be playing a key role in the control of gene expression in mammalian cells. The genes involved include tumor suppressor genes, genes that suppress metastasis and angiogenesis, and genes that repair DNA suggesting that epigenetics plays an important role in tumorigenesis. It becomes clear that methylation is a basic, vital feature/mechanism in mammalian cells. It is involved in hereditary and somatic cancers, hereditary and somatic diseases, apoptosis, replication, etc. It is suggested that it can be used for diagnostic, prognostic, prediction and even for direct treatment of cancer. Based on the growing understanding of the roles of DNA methylation, several new methodologies were developed to make a genome-wide search for changes in DNA methylation: restriction landmark genomic scanning (RLGS), methylation-sensitive-representational difference analysis (MS-RDA), methylation-specific AP-PCR (MS-AP-PCR), methyl-CpG binding domain column/segregation of partly melted molecules (MBD/SPM) and CpG islands microarrays (CGI). Although each of them has its own advantages none of them is suited for large-scale screening as all are rather inefficient, complicated or technically challenging; they can be used only for testing a few samples. For example, after analysis of 1000 clones isolated using MBD/SPM, nine DNA fragments were identified as CpG islands and only one was specifically methylated in tumor DNA. Maximum resolution of RLGS is 1000–2000 NotI boundary clones (i.e. appr. 700–800 NotI sites) and we have isolated more than 20000 (Kutsenko *et al*., 2002). Identification and quantification of microbial species in their various habitats is very important for understanding and dealing with different aspects of human and animal health and disease. For example, identification of pathogenic bacteria in food, soil or in air can prevent epidemics. Not very much is known about the human normal microflora. The human intestinal tract harbors a densely populated, active and complex bacterial ecosystem. The number of microbial cells in the colon is estimated to be 10–100 times larger than the number of eukaryotic cells in the human entire body and weighs more than one kg. Microscopic investigations demonstrated that many different species of microorganisms live in our digestive system, but at least 85 % are unknown, mainly because they cannot easily be cultured *in vitro*. At the same time many studies have shown that the composition of the gut microflora plays a very significant role for human health. Many intestinal bacteria are known to provide molecules that the host itself cannot manufacture or degrade from nutritional compounds. Thus, these organisms are clearly of survival value, they are true symbionts of the host. At the same time, several major diseases are believed to have the gut flora as the main potential source of pathogenesis (allergy, IBD, Crohn disease, cancer, etc.).Identification of bacterial species and strains relies heavily on culture techniques. However, in a complex bacterial population, rapidly growing bacteria would overgrow, making quantification and identification of slow or non-growing bacteria impossible. Techniques still have to be developed even to culture a representative selection of the microorganisms. Consequently, the picture of the intestinal flora has been biased in favor of the more easily cultured bacteria. There are some methods available to analyze complex microbial mixtures, e.g. by enzyme analysis which requires growth of colonies outside the body, or analysis of the fatty acids composition in stools, both of which give crude and indirect indications of the composition of the normal flora. The limitations of such techniques are obvious.The application of culture-independent techniques based on molecular biology methods can overcome some shortcomings of conventional cultivation methods. In recent years the approach based on PCR amplification of 16S rRNA genes has become both popular and very useful. One modification of the approach utilized fingerprinting of all the species in the gut using, for instance, denaturing gradient gel electrophoresis (DGGE) with PCR amplified fragments of 16S rRNA genes. In another application, PCR amplified fragments of 16S rRNA genes were directly cloned and sequenced. These studies provided important information, however intrinsic disadvantages of the approach limit its application. The problem is that 16S rRNA genes are highly conserved and therefore the same sequenced fragment sometimes can represent different species. It is also difficult to adapt for quantification. Moreover, in fingerprinting experiments similar fragments may represent different species and yet different fragments may also represent the same species.

Microarray technology using immobilized DNA has opened up new possibilities in molecular biology of eukaryotes and prokaryotes. This approach was also applied to the studies of the bacterial composition of the microflora and identification of specific microbial species. However, the microarrays based on 16S rRNA genes suffer from the same problems as sequencing/ fingerprinting methods and species-specific microarrays based on PCR amplification of specific DNA/gene fragments can only be used for identification of a limited number of microorganisms.

Methods

We suggested to use for the analysis of genomes and genome mixtures new methods: NotI passporting (tags) and NotI microarrays that were described in detail in (Li *et al*., 2002; Zabarovska *et al*., 2003; Zabarovsky *et al*., 2003). Important to note that for this analysis we use only some specific fragments of the genomes (NotI representations). Thus we do not aim to sequence all genomes or study all genes. We append special signatures for a particular organism/genes and analyze these signatures in different samples. Using the same idea (short sequence tags) we have also developed a new approach to genome mapping and sequencing based on slalom libraries

(Zabarovska *et al*., 2002). The concept represents alternative approaches to the construction of linking and jumping libraries, and involves the construction of "slalom libraries". The pilot experiments (Zabarovska *et al*., 2002) demonstrated the feasibility of the approach, and showed that the efficiency (cost-effectiveness and speed) of existing mapping/sequencing methods can be improved at least 5- to 10-fold. Furthermore, since the efficiency of contig assembly in the slalom approach is virtually independent of sequence read length, even short sequences, as produced by rapid high-throughput sequencing techniques suffice to complete a physical map and sequence scan of a small genome. Combination of these new sequencing techniques with slalom approach increase the power of the method 10–50 times more.

Results and Discussion

NotI microarrays for genome wide screening cancer cells. The fundamental problems for genome wide screening using NotI clones are: (i) the size and complexity of the human genome; (ii) the number of repeat sequences; and (iii) the comparatively small size of the inserts in NotI clones (on average 6–8 kb). To solve this problem, the special primers were designed and special procedure was developed to amplify only regions surrounding NotI sites, so called NotI representation (NR). Other DNA fragments were not amplified. We suggested to use for genome screening NotI microarrays in combination with this new method for labeling genomic DNA where only sequences surrounding NotI sites are labeled. A pilot experiment using NR probes demonstrated the power of the method, and NotI clones deleted in cancer lines, renal and breast biopsies were found. Experiments demonstrated that sensitivity of the method is enough to detect difference 1:2 (e.g. man from woman with the use of X chromosome specific NotI clones). Important to mention that in these experiments polymorphic and methylated NotI clones were also successfully detected. Our estimation is that human genome contains 10.000–15.000 NotI sites and 5.000–9.000 of them are unmethylated in a particular cell. Thus screening with NotI microarrays will be equivalent to screening using 3.000–4.000 gene associated single nucleotide polymorphisms (SNP). NotI microarrays give additional information to the deletion mapping: they can be used for gene expression profiling and methylation studies.

Moreover, NotI microarrays are the only existing microarrays giving the opportunity to detect methylation and copy number changes simultaneously or differentially. There is no reason why NotI microarrays cannot be used to study histone modifications and we are currently performing theses experiments. NotI microarrays have another strong advantage compared to cDNA microarrays. There is no standard for comparing expression profiles. RNA is not a stable molecule and physiological conditions of the cancer cells and normal cells can be rather different, moreover such conditions vary very significantly during the short time period, depending on many different factors, e.g. temperature, day time, psychological status, medicine treatment, etc. On the other hand, with expression microarrays it is practically impossible to find the first events and first genetical lesions that leads to the development of cancer. This is not a problem for NotI microarrays as genetic lesions (for example deletions) are irreversible, epigenetic changes (e.g. methylation) are not so temporal and the normal genomic DNA is an perfect standard for comparisons. These features are rather important for different studies like diagnostics, prognosis, prediction etc.

NotI microarrays for analysis of complex microbial systems. In a pilot experiment we have produced NotI microarrays from gram-positive and gram-negative bacteria and have shown that even closely related *E. coli* strains can be easily discriminated using this technique. For example, two *E. coli* strains, K12 and R2, differ in less than 0.1 % in their 16S rRNA sequences and thus the 16S rRNA sequence would not easily discriminate between these strains. However, these strains showed distinctly different hybridization patterns with NotI microarrays.The same technique can be adapted to other restriction enzymes as well. This type of microarray opens the possibility not only for studies of the normal flora of the gut but also for any problem where quantitative and qualitative analysis of microbial (or large viral) genomes is needed.

NotI passporting (generation of NotI tags) for analysis of complex microbial systems. We demonstrated that these tags comprising 19 bp of sequence information could be successfully generated using DNA isolated from intestinal or faecal samples. NotI passports allow the discrimination between closely related bacterial species and even strains. This procedure for generating restriction site tagged sequences (RSTS) is called passporting and can be adapted to any other rare cutting restriction enzyme. A comparison of 1 312 tags from available sequenced *E. coli* genomes, generated with the NotI, PmeI and SbfI restriction enzymes, revealed only 219 tags that were not unique. None of these tags matched human or rodent sequences. Therefore the approach allows analysis of complex microbial mixtures such as in human gut and identification with high accuracy of a particular bacterial strain on a quantitative and qualitative basis. Among all tags from all sequenced bacteria 97% are species-specific.

References

Li J. *et al.* NotI subtraction and NotI-specific microarrays to detect copy number and methylation changes in whole genomes // Proc. Natl Acad. Sci. USA. 2002. V. 99. P. 10724–9.

Kutsenko A. *et al. Not*I flanking sequences: a tool for gene discovery and verification of the human genome // Nucleic Acids Res. 2002. V. 30. P. 3163–3170.

Zabarovska V. *et al.* A new approach to genome mapping and sequencing: slalom libraries // Nucleic Acids Res. 2002. V. 30 (e6). P. 1–8.

Zabarovska V. *et al.* NotI passporting to identify species composition of complex microbial systems // Nucleic Acids Res. 2003. V. 31 (e5). P. 1–10.

Zabarovsky E. *et al.* Restriction site tagged microarrays (RST): a novel technique to identify the species composition of complex microbial systems // Nucleic Acids Res. 2003. V. 31 (e95). P. 1–8.

MODELING OF NEURO-ENDOCRINE-IMMUNE NETWORK VIA SUBJECT ORIENTED LITERATURE MINING

*Zhang C., Li S.**

Institute of Bioinformatics, MOE Key Laboratory of Bioinformatics, Tsinghua University, Beijing, P. R. China
* Corresponding author: e-mail: shaoli@tsinghua.edu.cn

Keywords: *model, neuro-endocrine-immune, network, subject oriented literature mining*

Summary

Motivation: To explore the functional communication and regulation of the neuro- endocrine-immune (NEI) network and its relationship with autoimmunity diseases via automated literature mining (LM).

Results: We applied subject oriented literature mining (SOLM) to the NEI literature and more than 14,000 PubMed abstracts for NEI are indexed automatically. Relating protein, gene, or small molecules (PGSM) interactions are extracted to model the NEI network, which is a holistic view of current dicoveries in this field and hopefully contains implications for future research. The characteristics of both the entire network and four sub-networks of the different NEI based pathologic changes, are evaluated and compared, and partly validated by our experimental results.

Introduction

The NEI network (Basedovsky, Sorkin, 1977) plays a pivotal role in modulating host homeostasis and optimizing health naturally. However, the lack of an integrated method is still a nodus for understanding the functional structure and systemic regulation of NEI. The system of neuroendocrine regulation of inflammatory, immune responses and disease is extremely complicated, which is beyond the scope of any individual researcher.

The rich resource of published biomedical literature motivated synergic efforts in the past decade to extract useful information from the literature, store itin a database with a computer-readable format, and easily make it available to researchers easily by manual curation, such as DIP, BIND and KEGG , or automated LM.

Up to now, most of the LM research for extracting PGSM interactions focuses on developing an automated system for a very general purpose, say, constructing an interaction database (Jenssen *et al*., 2001). However, PGSMs may have different interactions in different tissues or under different conditions. For example, IL-6, an important immunologic mediator, has 1177 neighbor factors, each of which is co-cited with IL-6 in at least 27 abstracts according to the report of PubGene in March, 2004 (http://www.pubgene.org). Obviously, for most biologists who are interested in the PGSM interections in the context of a speific topic or subject, it is very difficult to examine the results and select information they want. As far as we know, there are still a few reports, aiming at extracting PGSM interactions of a specific subject, such as NEI, from all the available literature. In this paper, subject oriented literature mining (SOLM) is applied to extract knowledge specific toNEI by incorporating prior knowledge from medical scientists and biologists in a straightforward manner.

Method

Our SOLM system is based on co-citation, sharing the assumption with several existing LM systems that if two PGSMs are co-cited in the same text unit (such as abstract, sentence and phrase), there should be an underlying biological relationship between them. To find co-citation, a pool of articles, for example, abstracts from PubMed, and a dictionary containing PGSM terms and their synonyms are required. In the process of selecting candicate articles and/or PGSM dictionary, prior knowledge can be conveniently incorporated. Our rationale is that this prior knowledge can help us to find interaction information particularly reflecting the priori, enen if limitted.

PGSM dictionary. As a first step, a biologist manually reads dozens of important NEI articles published in recent years and summarizes a list of PGSM terms and synonyms, which contain 129 PGSMs, corresponding to 271 names to be indexed. 275 interactions corresponding to 89 PGSMs are also extracted for comparison.

Preparation of articles. As listed in Table 1, 10 keywords closely related to NEI are submitted to PubMed and 14442 abstracts are downloaded (Dec 28, 2003).

Table 1. Preparation of articles. 14442 PubMed abstracts are downloaded

Keywords	Abstracts	Keywords	Abstracts
Neuro-endocrine-immune	488	Psychoneuroimmunology	721
Immune-neuroendocrine	71	Neuroimmunomodulation	1511
Neuroimmunoendocrinology	15	Hypothalamic-pituitary	7543
Neuroimmunology	1222	NF-kappaB pathway	2766
Neuroendocrinology	471	Pineal-immune	4

Text unit for co-citation. Co-citation can be calculated according to different text units, for example, full abstracts, constituent sentences, single sentences and phrases, from large to small. A pair of PGSMs co-cited (or biologically interacting with each other) in a larger text unit is not necessarily co-cited in a smaller one. Generally speaking the larger the text unit from which the co-citation is derived, the lower the precision ($\mathrm{p} = \mathrm{TP}/(TP + FP)$) but the higher the recall ($\mathrm{r} = \mathrm{TP}/(TP + FN)$), and vice versa. (Ding *et al.*, 2002) systematically studied the impact of different text units upon precision, recall and effectiveness ($\mathrm{e} = 2\mathrm{pr}/(p + r)$). Sentence as a text unit is found to make the best trade off between precision and recall with the highest effectiveness (TP: the number of true positives; FP: the number of false positives; FN: the number of false negatives). Accordingly, we regard two PGSMs as co-cited if they co-occur in the same sentence.

Normalization of co-citation. The co-citation number of an interaction is related to two factors: the importance of the interaction, and the discovered time of the interaction. On the one hand, important discoveries will be cited by more followers. On the other hand, earlier discoveries also tend to have a greater co-citation number. To highlight newly-discovered interactions, which may have more important implications for future research, normalization is necessary. Denote the co-citation matrix as $C = (c_{ij})_n$, where C_{ij} is the co-citation number between PGSM *i* and PGSM *j*, and *n* is the number of PGSM (*n*=114 in our case). The normalized co-citation is derived as follows

$$C_{norm} = (c'_{ij})_n \text{, where } c'_{ij} = \frac{2c_{ij}}{\sum_{k=1}^{n} c_{ik} + \sum_{k=1}^{n} c_{jk}} . \tag{1}$$

Obviously, $c'_{ij} \in [0, 1]$. If all the occurrences of PGSM i are accompanied by PGSM j and all the occurrences of PGSM j are also accompanied by PGSM i, $c'_{ij} = 1$. The normalized co-citation reflects the significance of interaction between PGSM *i* and PGSM *j* among all the co-citations of the two PGSMs with the other PGSMs.

Comparison of different networks. In a network, the topology of a PGSM i can be represented by

$v_i = [c_{i1}, c_{i2}, K, c_{in}]^T$,where $n = 114$. Thus, the difference of topology of PGSM in two networks is $d_i = \|v_i(1) - v_i(2)\|$. (2)

When we compare NEI sub-topic networks with the general NEI network, those PGSMs corresponding to a specific sub-topic can be highlighted with large topological change.

Results and Discussion

Among 14442 abstracts, 25104 co-citations are matched, corresponding to 1519 interactions between 114 factors of 129 PGSMs. Other 10 PGSMs are also matched in abstracts without co-citation, and they are excluded fromour further analysis. The distribution of co-citation number per PGSM follows power-law distribution (Fig. 1), which means PGSM interaction network is a scale free network, a property shared by many other networks. Compared with the manual work, 146 interactions of 275 (53 %) are included in our SOLM results, which is a comparable performance with others (Jenssen *et al.*, 2001).

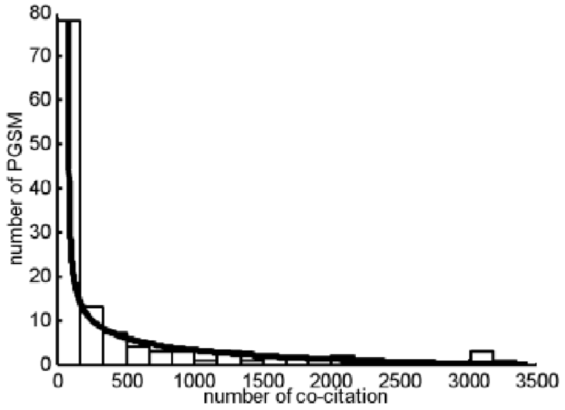


Fig. 1. The distribution of the co-citation number per PGSM follows the scale free power-law distribution. The solid line is the result of power-law curve-fitting.

In the model of the NEI network, 114 factors and their interactions are structured into two clusters of endocrine hormones and immunity cytokines that mediate the NEI interface. Fig. 2 shows a visually clarified sub-network, which is found to be biologically meaningful.

Furthermore, we selected subsets from the NEI literature for the sub-topics of "rheumatoid arthritis", "systemic lupus erythematosus", "cold" and "heat", respectively, and followed processing steps described in the sub-sections above, which are summarized in Table 2.

This study, with its results partially being validated by our experiment (Li *et al.*, 2004), may give an ordered list of PGSMs that shows strong implicit (indirect) interactions for the regulation of the NEI system and its relations to autoimmune diseases. As a future direction, the effectiveness of the results will be further validated by microarray data and protein interaction databases. The principal pathway and the key factors in the model will be subject to clinical observation for the verification of the network.

Table 2. Comparison of different sub-networks

Sub-topics	Keywords	Abstracts	PGSMs	Interactions	Co-citations
RA	"rheumatoid arthritis"	159	57	185	408
SLE	"systemic lupus erythematosus" or "SLE"	40	33	64	92
Cold	"cold"	121	38	74	285
Heat	"heat"	128	41	74	165

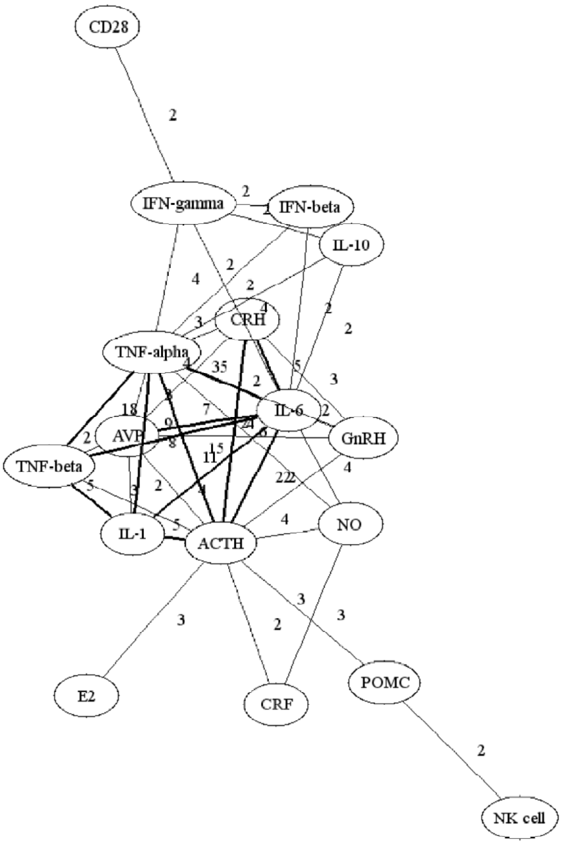


Fig. 2. Detailed sub-network of the combination of IL-6, ACTH and TNF-α in NEI. With IL-6, ACTH, TNF-α and their synonyms as keywords, 24 abstracts and 23 PGSMs are matched, 69 interactions and 253 co-citations are presented.

Acknowledgements

The work was supported by NSFC (No. 30200365) and the Key Project of the Chinese MOE (No.104009). We are grateful to Lijiang Wu and Xi Ma for helpful discussions.

References

Basedovsky H.O., Sorkin E. Network of immune-neuroendocrine interactions // Clin. Exp. Immunol. 1977. V. 27. P. 1–12.

Ding J., Berleant D., Nettleton D. *et al.* Mining MEDLINE: abstracts, sentences, or phrases? // PacSymp. Biocomput. 2002. V. 7. P. 326–337.

Jenssen T.K., Laegreid A., Komorowski J. *et al.* A literature network of human genes for high-throughput analysis of gene expression // Nat. Genet. 2001. V. 28. P. 21–28.

Li S., Lu A.P., Li B. *et al.* Circadian rhythms on HPA axis hormones and cytokines of collagen induced arthritis in rats // J. Autoimmun. 2004. V. 21. (In press).

BGRS
2004

COMPUTATIONAL EVOLUTIONARY BIOLOGY

EVOLUTION OF GENE STRUCTURE IN EUKARYOTIC GENOMES

Babenko V.N.[1, 2]****, Sverdlov A.***[1]***, Rogozin I.B.***[1, 2]***, Koonin E.V.***[1]

[1] National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD, USA; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: V.N. Babenko, NCBI/NLM/NIH, 8600 Rockville Pike, Bldg. 38A, Bethesda, MD 20894, USA; e-mail: babenko@ncbi.nlm.nih.gov, Fax: (301)480-4637

Keywords: *spliceosomal intron, evolutionary conservation, reverse transcription, homologous recombination*

Summary

Motivation: In spite of numerous computational studies, origins and mechanisms of evolution of eukaryotic spliceosomal introns remain mysterious. We approached these problems from a comparative-genomic standpoint, i.e., by comparing homologous gene structures in 7 eukaryotic lineages representing three kingdoms and using the parsimony principle to reconstruct evolutionary scenarios. This allowed us to identify previously unnoticed features of splice-site sequences and gene organization.
Results: In an attempt to gain insight into the dynamics of intron evolution in eukaryotic protein-coding genes, the distributions of old introns, which are conserved between distant phylogenetic lineages, and new, lineage-specific introns along the gene length were examined. A significant excess of old introns in 5'-regions of genes was detected. New introns, when analyzed in bulk, showed a nearly flat distribution from the 5'- to the 3'-end. However, analysis of new intron distributions in individual genomes revealed notable lineage-specific features. While in intron-poor genomes, particularly yeast *Schizosaccharomyces pombe*, the 5'-portions of genes contain a significantly greater number of new introns than the 3'-portions, the intron-rich genomes of humans and Arabidopsis show the opposite trend. These observations seem to be compatible with the view that introns are both lost and inserted in 3'-terminal portions of genes more often than in 5'-portions. Over-representation of 3'-terminal sequences among cDNAs that mediate intron loss appears to be the most likely explanation for the apparent preferential loss of introns in the distal parts of genes. Preferential insertion of introns in the 3'-portions suggests that introns might be inserted via a reverse-transcription-mediated pathway similar to that implicated in intron loss.

Introduction

Protein-coding genes of multicellular eukaryotes typically contain multiple introns; these introns are spliced from pre-mRNA by spliceosomes, complex molecular machines that read a conserved nucleotide pattern, which signals where splicing is to take place. Several studies on individual genes and, subsequently, on genome scale have shown that, although a substantial fraction of intron positions is conserved through hundreds of millions and even billions of years of eukaryotic evolution, numerous other introns are lost and gained (Fedorov *et al*., 2002; Rogozin *et al*., 2003).

It has been also noticed that distributions of intron positions over the length of the coding region differ substantially in different organisms. In particular, in intron-poor genomes of single-cell eukaryotes, introns are strongly over-represented in the 5'-portions of genes, whereas, in intron-rich multicellular organisms, the distribution is closer to uniformity (Sakurai *et al*., 2002; Mourier, Jeffares, 2003). A mechanistic explanation for this trend suggests that introns are preferentially lost from the 3'-portion of a gene due to the over-representation of 5'-truncated transcripts among the cDNAs that are produced by reverse transcription and are thought to mediate intron loss via homologous recombination (Fink, 1987; Derr, 1998). However, selectionist interpretations of the observed distributions of introns, such that introns located in the 5' part of a gene might be more tightly integrated into one or more of intron-mediated functions, have been proposed as well (Mourier, Jeffares, 2003).

Here we expand our previous analysis that allowed classification of introns in conserved eukaryotic genes into old ones, which are shared by two or more distant lineages, and relatively new, lineage-specific ones. We compare the distributions of old and new introns along the length of eukaryotic genes and show that while ancient introns are substantially over-represented in the 5'-portions of the genes in all sequenced eukaryotic genomes, new introns are distributed much more uniformly. Moreover, in the most intron-rich genomes, new introns are over-represented in the 3'-portions of the genes. These results can be interpreted as an indication that both loss and insertion of introns occur preferentially in the 3'-regions of genes, which suggests a reverse-transcription-mediated mechanism for both processes.

Materials and Methods

Old and new introns were identified by analysis of 684 clusters of eukaryotic orthologous groups (KOGs), each of which included orthologous genes from 8 eukaryotic species with (nearly) completely sequenced genomes: *Homo sapiens (Hs), Caenorhabditis elegans (Ce), Drosophila melanogaster (Dm), Saccharomyces cerevisiae (Sc), Schizosaccharomyces pombe (Sp), Arabidopsis thaliana (At), Anopheles gambiae (Ag),* and *Plasmodium falciparum (Pf)* (Rogozin *et al.*, 2003). Intron positions were tallied from the feature tables of complete genome annotations in the GenBank database. Alignments of protein sequences of KOG members were constructed using the MAP program and converted back to nucleotide sequence alignments as previously described (Rogozin *et al.*, 2003). Introns that occupied the exact same position in aligned amino acid sequences of KOG members were considered orthologous. Introns were partitioned into the old and new sets according to their conservation or lack thereof among distinct phylogenetic lineages. Specifically, within the analyzed set of genomes, introns that were conserved in the fly and the mosquito were considered new; in all other cases, introns conserved in two or more species were classified as old (ancient).

Results and Discussion

A notable difference was seen in the distributions of old and new introns along the length of the gene: when the data from all analyzed genomes were pooled together, the density of ancient introns markedly dropped from 5'-end to 3'-end of the coding sequence (CDS), whereas new introns showed a nearly uniform distribution across the entire length of the CDS (Table). The trend for enrichment of old introns in the 5'-part of genes held for each of the analyzed genomes and was statistically significant (P < 0.05) in all cases (Table).

Table. Comparison of intron abundance in the 5' and 3' halves of coding sequences of eukaryotic genes

Species	Number of old introns		Number of new introns		P_{old}	P_{new}	P_{Fisher}
	5'-region	3'-region	5'-region	3' -region			
At	598	510	1637	1847	0.005	0.0002	0.00003
Ag	430	350	240	216	0.002	0.141	0.214
Ce	487	381	947	874	0.0002	0.046	0.025
Dm	440	368	277	249	0.006	0.119	0.279
Hs	1045	880	1797	1860	0.00009	0.152	0.0001
Pf	90	64	254	236	0.02	0.221	0.09
Sp	191	122	165	91	0.00006	0.000002	0.225

For statistical calculations, the first and last of the 10 bins were disregarded to eliminate potential artifacts caused by the fact that the terminal exons often extend into untranslated regions. Thus, bins 2-5 were combined into the 5'-region of the coding sequence, whereas bins 6–9 comprised the 3'-region. P_{old} is the *P*-value corresponding to the null hypothesis that old introns are equally distributed between the 5'-region and the 3'-region.and P_{new} is the probability of the same null hypothesis for the new introns. The probabilities were calculated using the binomial test; P_{fisher} is the probability that the distributions of of old and new introns along the length of the coding sequences are the same as calculated using Fisher's exact test.

The distributions of new introns showed opposite trends for intron-rich and intron-poor genomes. In intron-poor, unicellular organisms, *S. pombe* and *P. falciparum*, and in the relatively intron-poor animals, the nematode and the two insects, the density of new introns also dropped from 5'-end to 3'-end, although the trend was typically less pronounced than that for old introns and was statistically significant only for *P. falciparum* and *C. elegans*. In contrast, in the intron-rich species, *A. thaliana* and *H. sapiens*, there was an increase in the density of new introns from the 5'-end to the 3'-end of the coding region; the excess of new introns in the 3'-terminal portions of genes was statistically significant in the case of Arabidopsis. A comparison of the distributions of old and new introns along the gene length using Fisher's exact test showed that the difference was highly statistically significant in the case of Arabidopsis and humans, and moderately significant in the case of *C. elegans* (Table). An additional comparison of the slopes of the regression lines for old and new introns showed that, in all analyzed species (with the notable exception of Plasmodium), the slope was consistently greater for old introns and, as a whole, the difference was statistically significant (results not shown).

New introns, when analyzed in bulk, showed a nearly uniform distribution. Moreover, in intron-rich species, the 3'-portions of genes were enriched for new introns. The intron-rich genomes represent evolutionary lineages in which intron gain appears to dominate over intron loss (Rogozin *et al.*, 2003). Thus, over-representation of new introns in 3'-portions of genes from these species suggests that intron insertion might have the same bias as intron loss, i.e., introns could preferentially insert into the distal part of the coding region. Although the differences between the distributions of old and new introns were highly significant, the excess of new introns in 3'-portions of genes in intron-rich species was relatively slight. However, this potentially could reflect a steep gradient of intron insertion rates over the gene length. Indeed, the observed distribution of introns results from the balance between intron insertion and loss. For example, if the ratio of the rates of intron loss in the 3'- and 5'-halves of a gene is 2:1, the ratio of the rates of intron insertion should be the same to ensure a uniform distribution of introns.

The mechanism(s) of insertion and origin of new introns remain unknown. Direct attempts on identification of intron sources by comparison of intron sequences to other parts of the respective genomes failed to produce any substantial clues. Furthermore, comparative analysis of information content of splice signals of old and new introns suggested that, even assuming rapid decay of sequence similarity, old introns were unlikely to serve as a major source of new ones (Sverdlov *et al.*, 2003). The apparent preferential insertion of introns in the 3'-portions of genes suggests that one of the major mechanisms of introns gain might be a reverse-transcription-mediated pathway similar to that implicated in intron loss. One could speculate that this process involves duplication of a portion of the CDS during reverse transcription followed by homologous recombination, with one of the duplicates becoming an intron, sequence similarity between the intron and the adjacent exon is likely to decay too rapidly to be detectable. Thus, analysis of the non-uniform and distinct distributions of old and new introns in eukaryotic genes might provide clues to the poorly understood mechanics of intron evolution.

References

Derr L.K. The involvement of cellular recombination and repair genes in RNA-mediated recombination in Saccharomyces cerevisiae // Genetics. 1998. V. 148. P. 937–45.

Fedorov A., Merican A.F., Gilbert W. Large-scale comparison of intron positions among animal, plant, and fungal genes // Proc. Natl Acad. Sci. USA. 2002. V. 99. P. 16128–33.

Fink G.R. Pseudogenes in yeast? // Cell. 1987. V. 49. P. 5–6.

Mourier T., Jeffares D.C. Eukaryotic intron loss // Science. 2003. V. 300. P. 1393.

Rogozin I.B., Wolf Y.I., Sorokin A.V., Mirkin B.G., Koonin E.V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution // Curr Biol. 2003. V. 13. P. 1512–7.

Sakurai A., Fujimori S., Kochiwa H., Kitamura-Abe S., Washio T., Saito R., Carninci P., Hayashizaki Y., Tomita M. On biased distribution of introns in various eukaryotes // Gene. 2002. V. 300. P. 89–95.

Sverdlov A.V., Rogozin I.B., Babenko V.N., Koonin E.V. Evidence of splice signal migration from exon to intron during intron evolution // Curr. Biol. 2003. V. 13. P. 2170–2174.

CHANGE IN CpG CONTEXT IS A LEADING CAUSE OF CORRELATION BETWEEN RATES OF NON-SYNONYMOUS AND SYNONYMOUS SUBSTITUTIONS IN RODENTS

Bazykin G.A.[1]*, ***Ogurtsov A.Y.***[2], ***Kondrashov A.S.***[2]

[1] Dept. of Ecology and Evolutionary Biology, Princeton University, Princeton, NJ 08540, USA; [2] National Center for Biotechnology Information, NIH, Bethesda, Maryland 20894, USA
* Corresponding author: e-mail: gbazykin@princeton.edu

Keywords: *evolution, point substitution rate, mutation bias, CpG deamination, dinucleotides*

Summary

Motivation: Correlation between the rates of synonymous (silent) and non-synonymous (amino acid-changing) nucleotide substitutions in genes is a wide-spread and yet unexplained genome-level phenomenon which is in disagreement with the neutral theory of molecular evolution (Kimura, 1983). In mammals, this correlation can be caused by mutational dependence of the point mutation events.

Results: Comparison of 7,732 alignments of mouse and rat genes confirms the previously observed correlation between rates of substitutions in non-degenerate (K_A) and four-fold degenerate (K_4) nucleotide sites. In rodents, this correlation is primarily caused by tandem substitutions and, in particular, by CpG mutation bias leading to doublet nucleotide substitutions. The nature of correlation between K_A and K_4 in seven pairs of prokaryotic genomes is unclear.

Introduction

Synonymous (silent) nucleotide sites are often assumed to evolve “neutrally” and therefore are frequently used as a measure of non-synonymous substitutions. This assumption, however, conflicts with the well-described phenomenon of variation of rates of synonymous substitutions across the genome and, in particular, of correlation between rates of non-synonymous and synonymous substitutions. Selection for translation efficiency (Chamary, Hurst, 2004) or RNA structure (Smith, Hurst, 1999) acting on silent sites were suggested as possible explanations, as well as methodological biases in distance estimation (Bielawski *et al*., 2000).

It has been claimed that the correlation of rates of synonymous and non-synonymous substitutions is dependent upon the particular method used for estimation of substitution rates (Bielawski *et al*., 2000). Therefore, to reveal the leading cause of this correlation, it is preferable to use closely related species. At low evolutionary distances, substitutional saturation is negligible, and different methods of estimation of divergence converge.

Methods

Mouse and rat coding sequences were obtained from version 30 of the mouse genome (Mouse Genome Sequencing Consortium, 2002) and version 2 of the rat genome (Rat Genome Project Sequencing Consortium, 2004) from NCBI. Orthologs were identified according to the two-directional best-hit approach using protein BLAST (Altschul *et al*., 1997). Alignments of the amino acid sequences for each pair of the orthologs was made using ClustalW (Thompson *et al*., 1994) and reverse transcribed to get the nucleotide alignments. Rates of nucleotide substitutions in different groups of sites were obtained using a PERL script available from the authors. All suitable triplets of bacterial genomes were obtained from the NCBI Entrez database and processed analogously. Genes with doublets removed are those in which adjacent nucleotide sites were excluded from analysis if both carried substitutions. A substitution at site 1 of the doublet was assumed to change the CpG context of the following site 2 when one of the species carried "C" at site 1 and the other species carried some other nucleotide. A substitution at site 2 of the doublet was assumed to change the CpG context of the preceding site 1 when one of the species carried "G" at site 2 and the other species carried some other nucleotide.

Outliers can have a profound effect on the value of correlation coefficient. In order to ensure that only high-quality (unambiguous) alignments are included in the analysis, we excluded all genes with divergences in non-degenerate sites exceeding 1.5 average amino-acid divergences between corresponding species, and divergences in 4-fold-degenerate sites exceeding 10 average amino-acid divergences (therefore the abrupt left and top boundaries of region with data points at Fig. 1a). This approach is conservative in regard to determination of correlation.

Results and Discussion

Our data confirms the previously observed significant correlation between per gene substitutions rates in non-synonymous and synonymous nucleotide sites (Fig. 1a). This correlation, however, is primarily caused by doublet substitutions occurring in adjacent nucleotides. When sites with double substitutions were excluded from analysis, the magnitude of correlation was greatly reduced (Fig. 1b).

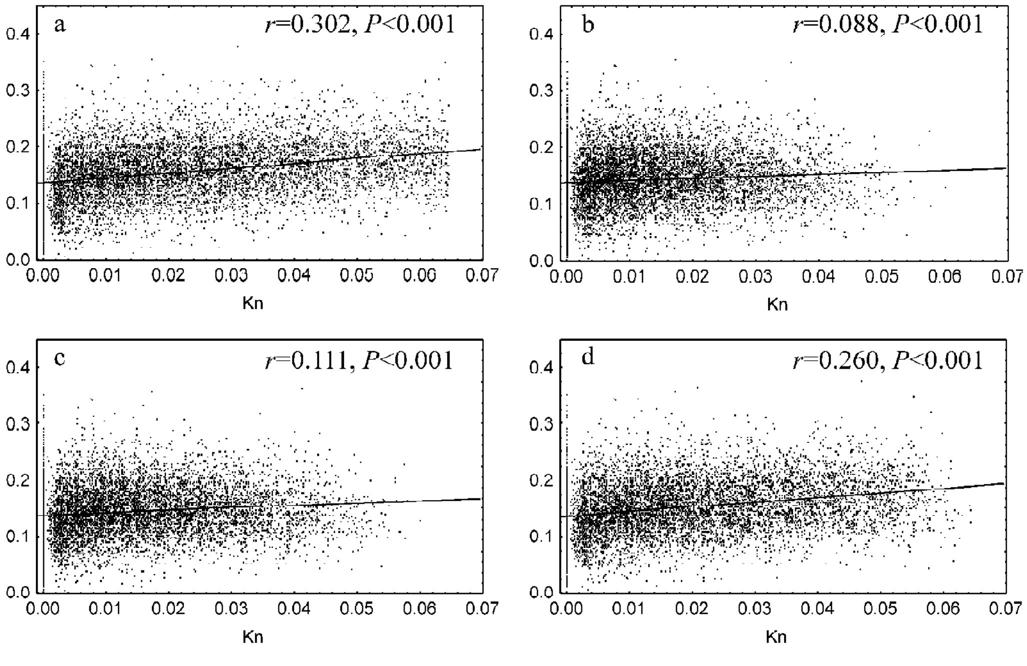


Fig. 1. The relationship between per gene divergences in non-degenerate (K_A) and four-fold degenerate (K_4) nucleotide sites between mouse and rat with all sites included into analysis (a), doublets removed (b), doublets with change in CpG context removed (c), and doublets without change in CpG context removed (d).

Correlation between substitutions in adjacent sites can arise if one mutational event simultaneously affects two successive nucleotides. However, such double substitutions are extremely rare (Kondrashov, 2003), and the observed effect has to be caused by separate point mutation events. Such correlation can also be due to selection on silent substitutions that restore codon bias following an amino acid change (Lipman, Wilbur, 1984).

The nature of correlation is revealed by consideration of the sites of adjacent substitutions in which on of the substitutions can affect the CpG context of the neighbouring nucleotide site. Removal of the subset of such sites is sufficient to achieve the strong reduction in correlation (Fig. 1c). Conversely, only a minor reduction in the correlation coefficient is achieved by removal of sites of neighbouring substitutions in which both substitutions leave the CpG context of the other one invariant (Fig. 1d).

The simplest explanation for correlation between K_A and K_4 that is consistent with these findings is interdependence of mutational events in adjacent nucleotides due to CpG deamination. CpG dinucleotide is hypermutable in vertebrates. If the first substitution (regardless of whether it occurs in a non-synonymous or synonymous site) creates the CpG dinucleotide, the second substitution at the adjacent nucleotide site is facilitated. This is expected to result in the observed pattern of substitutions coupling.

This explanation is further supported by analysis of seven pairs of closely related bacterial genomes. All the pairs of bacterial species indicated significant correlation between K_A and K_4 of various magnitude. However, removal of doublets and, in particular, of doublets involving change in CpG context did not lead to a profound decrease in correlation comparable with that observed in rodents. Therefore, some other factor has to be responsible for correlation between K_A and K_4 in prokaryotes.

Table. Correlation coefficients between divergences in non-degenerate and four-fold degenerate nucleotide sites in 8 pairs of genomes

	No. of genes	Fraction of amino acid differences[1]	All sites	Doublets removed[2]	Doublets with change in CpG context removed[2]	Doublets without change in CpG context removed[2]
Muridae	7.732	4.3 %	0.3	0.09	0.11	0.26
Bacillus	1.915	3.5 %	0.44	0.34	0.36	0.42
Bordetella	2.696	0.4 %	0.12	0.10	0.11	0.12
Ecoli	3.122	1.2 %	0.27	0.21	0.23	0.25
Salmonella	2.531	0.8 %	0.21	0.17	0.17	0.20
Staphylococcus	1.591	0.5 %	0.23	0.20	0.21	0.21
Streptococcus	1.065	0.7 %	0.28	0.20	0.22	0.26
Vibrio	579	0.7 %	0.29	0.25	0.26	0.28

The following pairs of genomes were analysed. Muridae: (*Rattus norvegicus*, *Mus musculus*); *Bacillus*: (*B. cereus* ATCC 14579, *B. anthracis* strain Ames); *Bordetella*: (*B. parapertussis*, *B. bronchiseptica* RB50); *Escherichia*: (*E. coli* O157:H7, *E. coli* K12); *Salmonella*: (*S. typhimurium* LT2, *S. enterica enterica* serovar Typhi Ty2); *Staphylococcus*: (*S. aureus aureus* Mu50, *S. aureus aureus* MW2); *Streptococcus*: (*S. pyogenes* M1 GAS, *S. pyogenes* MGAS315); *Vibrio*: (*V. vulnificus* YJ016, *V. vulnificus* CMCP6).

[1] Fraction of mismatches in alignments of orthologous proteins between genomes; [2] see Methods for details. All correlations were significant at $P < 0.05$.

An obvious next step would be to reveal the order of substitutions – whether the change of context in non-synonymous site facilitates the synonymous substation or vice versa. This can be achieved if a third orthologous gene from an outgroup species (e.g., human) is employed.

Acknowledgements

G.B. was supported by BWF graduate fellowship.

References

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. 1997. V. 25. P. 3389–3402.

Bielawski J.P., Dunn K.A., Yang Z. Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions // Genetics. 2000. V. 156. P. 1299–1308.

Chamary J.V., Hurst L.D. Similar rates but different modes of sequence evolution in introns and at exonic silent sites in rodents: evidence for selectively-driven codon usage // Mol. Biol. Evol. 2004 (advance online publication).

Kimura M. The Neutral Theory of Molecular Evolution. Cambridge Univ. Press, Cambridge. 1983.

Kondrashov A.S. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases // Hum Mutat. 2003. V. 21. P. 12–27.

Lipman D.J., Wilbur W.J. Interaction of silent and replacement changes in eukaryotic coding sequences // J. Mol. Evol. 1984. V. 21. P. 161–167.

Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome // Nature. 2002. V. 420. P. 520–562.

Rat Genome Sequencing Project Consortium. Genome sequence of the Brown Norway rat yields insights into mammalian evolution // Nature. 2004. V. 428. P. 493–521.

Smith N.G., Hurst L.D. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents // Genetics. 1999. V. 153. P. 1395–1402.

Thompson J.D., Higgins D.G., Gibson T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. V. 22**.** P. 4673–4680.

EVOLUTIONARY RELATIONSHIPS AND DISTRIBUTION OF NON-LTR RETROTRANSPOSONS IN EUKARYOTES

Beresikov E.[3]***, Novikova O.****[1]***, Makarevich I.***[2]***, Lashina V.***[1]***, Plasterk R.***[3]***, Blinov A.***[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Department of Agronomy and Plant Genetics, University of Minnesota, St. Paul, USA; [3] Hubrecht Laboratory, Netherlands Institute for Developmental Biology, Utrecht, The Netherlands

* Corresponding author: e-mail: novikova@bionet.nsc.ru

Keywords: *mobile elements, non-LTR retrotransposons, distribution, evolution, computer analysis, tree of life*

Resume

Non-LTR retrotransposons constitute a substantial part of some eukaryotic genomes and can play an important role in such evolutionary processes as genome rearrangement, gene formation and transcriptional pattering. Phylogenetic relationships between different families of non-LTR retrotransposons, based on sequences of reveres transcriptase, endonuclease and RNase H domains, are well-established, but no systematic study on the distribution of non-LTR elements among living organisms has been performed to date. We have screened 30 species, which represent all major eukaryotic phyla, for thepresence of 19 different families of non-LTR retrotransposons in their genomes using PCR with family-specific consensus-degenerate hybrid oligonucleotide primers. In addition, we have used our previously developed computational approach to search for non-LTR reverse transcriptase sequences in the data generated by 53 different genome sequencing projects. The combination of the two approaches resulted in the identification of 87 new non-LTR elements, belonging mostly to known families, in 44 species, thus doubling the number of species with characterized non-LTR retrotransposon content. The data obtained in this study allowed us to project the known phylogeny of non-LTR elements onto the Tree of Life, coupling the evolution of non-LTR retrotransposons with the evolution of their host genomes. Our findings support the principle steps of the previously proposed scenario for the evolution of non-LTR elements, such as order of acquisition of enzymatic domains and vertical mode of inheritance, but also provide additional resolution power to the current understanding of the evolution of non-LTR elements. Based on the structural, phylogenetic and distribution data, we propose a refined taxonomy of non-LTR retrotransposable elements.

Introduction

Retrotransposons are mobile genetic elements that propagate themselves by reverse transcription of an RNA intermediate. There are two major classes of retrotransposons, which differ structurally and mechanistically: LTR retrotransposons possess long terminal repeats (LTRs) and have a transposition mechanism similar to that of retroviruses, whereas non-long terminal repeat (non-LTR) retrotransposable elements do not have terminal repeats and utilize a simpler target-primed reverse transcription (TPRT) mechanism for their retrotransposition.

In TPRT, the element-encoded endonuclease cleaves the genomic DNA, the reverse transcriptase uses this break to prime reverse transcription from the element RNA, and the resulting cDNA copy is then integrated into the target site (Luan *et al.*, 1993). SINE (short interspersed nucleotide elements) elements use the non-LTR retrotransposon machinery for their transposition (Ogiwara *et al.*, 2002). Non-LTR retrotransposons have been found in all eukaryotes investigated to date and are the most abundant class of transposable elements. The copy number of the elements may vary from several copies per genome, as has been shown for some elements in *Drosophila melanogaster* (Berezikov *et al.*, 2000), to more than 800.000 copies (~20 % of the genome) for L1 elements in human (International Human Genome Sequencing Consortium, 2001). Transposition of non-LTR elements causes a hybrid disgenesis in *Drosophila* (Fawcett *et al*., 1986) and genetic diseases in human (Kazazian, Moran, 1998), and has been implicated in the emergence of pseudogenes (Esnault *et al*., 2000) and exon shuffling (Moran *et al*., 1999). Thus, non-LTR retrotransposons play important roles in the structural organization and evolution of the genomes they inhabit.

Phylogenetic analysis of non-LTR retrotransposons based on the reverse transcriptases domains allowed to distinguish 15 phylogenetic clades. Based on structural and phylogenetic features of different elements, Malik, Burke and Eickbush (1999) developed a scenario for evolution of non-LTR retrotransposons and demonstrated that non-LTR elements are inherited strictly by vertical transmission. Only a few cases of possible horizontal transfer of non-LTR retrotransposons have been suggested in literature (Kordis, Gubensek, 1999). According to the scenario of Malik, Burke and Eickbush, the most ancient clades of non-LTR retrotransposons (GENIE, CRE, R2, NeSL-1, and R4) contain only one ORF and show site-specific distribution in the genomes (Malik *et al*., 1999; Malik, Eickbush, 2000), which is provided by restriction-enzyme-like endonucleases (REL-endo) these elements encode. During further evolution of mobile elements, the REL-endo domain was substituted with an apurinic/apyrimidinic (AP) endonuclease acquired from the host cells. All younger clades (L1, RTE, Tad, R1, LOA, I, Jockey, CR1, Rex1, and L2) possess the AP endonuclease domain. The acquisition of the AP endonuclease resulted in losing target site specificity for all the elements (except the R1 clade and some elements from the L1 clade), and coincided with the origin of a second ORF in front of the RT-encoding ORF. Finally, elements of some clades obtained one more enzymatic domain in the second ORF – the RNase H domain.

It is clear that the presented distribution of the clades among living organisms reflects the extent of non-LTR retrotransposons investigation. Thus, representatives of nine out of fifteen known clades have been found in insects (mainly in dipteran species), where mobile elements were extensively characterized. A new approach for a broad investigation of the non-LTR retrotransposons distribution and evolution relies on a search for new transposable elements using the degenerate primers specific for the most conservative parts of the elements (Lovsin *et al*., 2001). Using this approach, detailed distribution of non-LTR retrotransposons has been analyzed for main taxa of the phylum Metazoa (Archipova, Meselson, 2000). Despite a large number of non-LTR retritransposon studies, no complete characterization of the distribution of these elements among different phyla in Eukaryotes has been made to date. Such characterization would provide new insights into the origin of non-LTR elements, their evolution and dispersal among living organisms, and also into the functional roles that non-LTR retrotransposons may play in multiple evolutionary processes, such as speciation, genome rearrangements, etc.

Materials and Methods

Total DNA isolation. Total DNA was isolated as described previously (Guryev *et al*., 2001).

PCR amplification and sequencing. Based on the comparison of 72 sequences of non-LTR retrotransposon reverse transcriptase domains, degenerate oligonucleotides primers were constructed using CODEHOPE software (Rose *et al*., 1998) to amplify a 500 bp region of reverse transcriptase. In total, eleven *sense* and eleven *anti-sense* degenerate primers were selected. Nineteen combinations of these primers were unique to nineteen selected families of non-LTR retrotransposons PCR amplification was performed using 0.1 μg of genomic DNA in 10- μl volume of 10 mM Tris-HCl (pH 8.9), 1 mM $(NH_4)_2SO_4$, 1.5 mM $MgCl_2$, 200 μM each of four dNTPs, 0.5 μM primers, and 2.5 units of *Taq* polymerase. After an initial denaturation step for 3 min at 94 °C, the PCR reactions were subjected to 30 cycles of amplification consisting of 30 sec denaturation at 94°C, 42 sec annealing at 52 °C, and 1 min extension at 72 °C. PCR results were assayed by agarose gel electrophoresis and PCR fragments of expected size were cloned into a pBlueScript (KS+) vector using standard procedures. The inserts were sequenced using DyeNamic ET chemistry on an ABI 3700 sequencer.

Sequence analysis. Search for non-LTR elements in publicly available genome sequencing data was performed as described previously (Berezikov *et al.*, 2000). The newly identified RT sequences were aligned to the previously established alignment of non-LTR reverse transcriptases (Malik *et al.*, 1999) using Clustal W software. The alignment used for phylogenetic tree construction is available as supplementary material. Phylogenetic trees were generated by Neighbor-Joining method using MEGA2 software package (Kumar *et al.*, 2001).

Results and Discussion

Screening of representative species by PCR with degenerate primers. For amplification of RT sequences from different organisms, we designed degenerate primers using the consensus-degenerate hybrid oligonucleotide primers approach (Rose *et al.*, 1998). These primers have a short degenerate 3' core region and a longer consensus 5' region, allowing amplification of distantly related sequences. Only 3 to 4 conserved amino acid residues in the analyzed group of sequences are essential for primer design. We have analyzed alignment of the RT sequences from 72 non-LTR elements (Malik *et al.*, 1999) and distinguished 19 families of retrotransposons based on the identity of amino acid motifs in the most conserved regions of the alignment. To investigate distribution of the families of non-LTR retrotransposons among living organisms, we selected thirty representative species that cover main eukaryotic taxa. The nineteen primer combinations were used to screen genomes of the thirty selected species for the presence of different families of non-LTR retrotransposons. The results of PCR were considered positive if a band of expected size (~500 bp) was observed. The PCR screening results were in good agreement with the known data of non-LTR retrotransposons distribution. To determine the nature of PCR fragments amplified by degenerate primers, we cloned and sequenced PCR products from one or several species for most of the non-LTR retrotransposon families. In total, 197 clones belonging to the different families were sequenced. Sequencing of the clones confirmed that degenerate PCR primers amplified the actual RT domains of non-LTR retrotransposons.

Computational screen for non-LTR reverse transcriptase sequences. Data generated by genome sequencing projects can provide definitive information on the structure and distribution of transposable elements among living organisms. We used our previously developed computational approach (Berezikov *et al.*, 2000) to screen data, generated by various sequencing projects, for non-LTR retrotransposons. The approach uses the profile hidden Markov model (HMM) software to find sequences matching the reverse transcriptase model and containing the motif F(Y)XDD, which is conserved among all reverse transcriptases. At this step, most of the potential reverse transcriptase sequences are identified. Next, BLAST analysis is performed to group the sequences by homology to reverse transcriptases of LTR or non-LTR elements, telomerases or retroviruses. Finally, redundancy of sequences in each group is removed and phylogenetic tree is constructed to estimate relationships between newly identified and known reverse transcriptases. We analyzed sequences of 53 eukaryotic organisms. Altogether, using our computational approach we identified 70 different non-LTR reverse transcriptase sequences in 35 species. Among these sequences, 29 directly corresponded to already known elements, whereas 41 sequences were new reverse tarnscriptases from 22 previously uncharacterized species. Most of the newly identified RTs clearly grouped with the known elements on a phylogenetic tree and covered all the clades of non-LTR retrotransposons.

Distribution of different non-LTR retrotransposon families among living organisms and their evolutionary relationships. The evolutionary scenario for non-LTR retrotransposons proposed by Malik *et al.* (1999) was based on a large but insect-biased dataset of non-LTR elements. Moreover, additional clades of retrotransposons have been described since the publication of their work. Our systematic PCR screening for non-LTR elements in representative species from major phylogenetic groups, as well as computational analysis of sequenced genomes, have substantially increased the number of species and elements in a dataset, thus allowing us to propose a refined model of the evolution and distribution of different families of non-LTR retrotransposons among eukaryotes. It should be noted that our model is principally the same as the model developed by Malik *et al.* (1999), and strongly relies on their finding that non-LTR elements evolve mainly through vertical transmission, and not through horizontal transfer. However, besides the phylogeny of the retrotransposon sequences themselves, our model also incorporates the phylogeny of host species, thus providing additional resolution power and a more integrative view for the evolution of non-LTR elements. Finally, we would like to present our version of the taxonomy of non-LTR retrotransposable elements that is based on all available data on the structural organization and evolutionary relationships of these elements. All non-LTR elements are divided into three groups according to the presence of one or two ORFs and AP and RNAseH domains. The first group of elements containing only one ORF includes 5 clades (GENIE, CRE, R4, NeSL-1, and R2) with one family in each clade. The second group of non-LTR retrotransposons containing two ORFs and the AP endonuclease domain also includes 5 clades (L1, RTE, CR1, Rex1, and Jockey) that together comprise 15 families. The last group containing two ORFs, AP endonuclease and RNaseH domains includes two clades, Tad and I, with three families in each of these clades. The taxonomy proposed here is most likely not the final version and will be changing with discoveries of new elements. However, the major division of non-LTR elements into clades will probably remain unchanged.

References

Arkhipova I., Meselson M. Transposable elements in sexual and ancient asexual taxa // Proc. Natl Acad. Sci. USA. 2000. V. 97. P. 14473–14477.

Berezikov E., Bucheton A., Busseau I. A search for reverse transcriptasecoding sequences reveals new non-LTR retrotransposons in the genome of *Drosophila melanogaster* // Genome Biol. 1:RESEARCH0012. 2000.

Esnault C., Maestre J., Heidmann T. Human LINE retrotransposons generate processed pseudogenes // Nat. Genet. 2000. V. 24. P. 363–367.

Fawcett D.H., Lister C.K., Kellett E., Finnegan D.J. Transposable elements controlling I-R hybrid dysgenesis in *D. melanogaster* are similar to mammalian LINEs // Cell. 1986. V. 47. P. 1007–1015.

Guryev V., Makarevitch I., Blinov A., Martin J. Phylogeny of the genus *Chironomus* (Diptera) inferred from DNA sequences of mitochondrial cytochrome b and cytochrome oxidase I. // Mol. Phyl. Evol. 2001. V. 19. P. 9–21.

Kazazian H.H., Moran J.V. The impact of L1 retrotransposons on the human genome // Nat. Genet. 1998. V. 19. P. 19–24.

Kordis D., Gubensek F. Molecular evolution of Bov-B LINEs in vertebrates // Gene. 1999. V. 238. P. 171–178.

Kumar S., Tamura K., Jakobsen I.B., Nei M. MEGA2: molecular evolutionary genetics analysis software // Bioinformatics. 2001. V. 17. P. 1244–1255.

Lovsin N., Gubensek F., Kordis D. Evolutionary dynamics in a novel L2 clade of non-LTR retrotransposons in Deuterostomia // Mol. Biol. Evol. 2001. V. 18. P. 2213–2224.

Luan D.D., Korman M.H., Jakubczak J.L., Eickbush T.H. Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition // Cell. 1993. V. 72. P. 595–605.

Malik H.S., Burke W.D., Eickbush T.H. The age and evolution of non-LTR retrotransposable elements // Mol. Biol. Evol. 1999. V. 16. P. 793–805.

Malik H.S., Eickbush T.H. NeSL-1, an ancient lineage of site-specific non-LTR retrotransposons from *Caenorhabditis elegans* // Genetics. 2000. V. 154. P. 193–203.

Moran J.V., Deberardinis R.J., Kazazian H.H. Jr. Exon shuffling by L1 retrotransposition // Science. 1999. V. 283. P. 1530–15304.

Ogiwara I., Miya M., Ohshima K., Okada N. V-SINEs: a new superfamily of vertebrate SINEs that are widespread in vertebrate genomes and retain a strongly conserved segment within each repetitive unit // Genome Res. 2002. V. 12. P. 316–324.

Rose T.M., Schultz E.R., Henikoff J.G., Pietrokovski S., Mccallum C.M., Henikoff S. Consensus-degenerate hybrid oligonucleotide primers for amplification of distantly related sequences // Nucleic Acids Res. 1998. V. 26. P. 1628–1635.

RELATIONSHIPS BETWEEN GENERAL CLASSIFICATION OF GENES' LATENT TRIPLET PERIODICITY AND THE UNIVERSAL PHYLOGENETIC TREE

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

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

Keywords: *latent triplet periodicity, evolution of genes, phylogenetic tree*

Resume

Motivation: The discussion of the ancient universal ancestor and probable components of its molecular genetic system, also of how the three main domain of life; Archceabacteria, Bacteria and eucaryotes, have arisen, is becoming very acute now. This is because huge enormous genetic material has been deciphered and also because numerous mathematical methods have been developed for its analysis. Investigation of the latent triplet periodicity of genes, and definition of the periodicity classes allow us to describe encoder "backbones" for all organisms. Here, we present a preliminary overview of gene systematisation according to the latent triplet periodicity classes, and a phylogenetic tree inferred from similarity of the genomes' encoder "backbones" for certain, archceabacteria, bacteria and eukaryotes.

Results: A search for the latent triplet periodicity was performed in the KEGG database, which contains the genes of various organisms. The proportional of the genes with latent periodicity amounts to at least 75 % for every organism. An algorithm of classification of the revealed periodicity was proposed. The relationships among organisms were determined on the basis of their gene distributions over the revealed classes of triplet periodicity (on the basis of the so-called encoder "backbones"). It was shown that the three major domains of life (Archaea, Bacteria and Eucaryotes, are differ by rhe content of the classes of their gene triplet periodicity.

Introduction

The Projects of sequence deciphering of the whole genomes provide a abundant of data for mathematical analysis of genome structure, functional content and evolutionary origin. The latent triplet periodicity probably arose as a result of encoding of amino acids by nucleotide triplets. The consecutive duplications of short DNA fragments have lead to creation of new coding sequence. The duplications of large coding sequences gave rise to many gene families (Ohno, 1970). Here, we search for the latent triplet periodicity in the sequences of known genes and classify the periodicity matrices aiming to outline a probable basic set of the original coding sequences of the ancient genome. We also attemped to establish the relations between the classes of triplet periodicity and their impact on genome structure of the organisms belonging to different taxons.

Methods and Algorithms

An previously proposed method of search for the latent periodicity in symbolic sequences (Korotkov *et al.*, 2003) was applied to search for the latent triplet periodicity in the genes of the KEGG-25 database (http://www.genome.ad.jp/kegg/). The search for such a periodicity in the 416.429 genes resulted in the 466.183 matrices of latent periodicity, which reflected the A, T, C, G nucleotides distribution over period positions in the sequences with the latent periodicity.

Algorithm of matrices classification

The following algorithm was applied to classify the matrices. Because the regions of the latent periodicity were of different lengths, the all the compared matrices were normalised at medium length of 1.002 nucleotides. Denote the period length as N (N=3), each matrix of the latent periodicity was represented as a vector of nucleotide quantities distributed over 4N ranks. The pair comparison was done between the vectors, as shown in Fig. 1. The lower index in Fig. 1 corresponds to the period position, the upper reflects a number of compared vector. Thus a matrix M1 was formed with the marginal quantities $X(i) = \Sigma_j M1(i,j)$, and $Y(j) = \Sigma_i M1(i,j)$, where $\Sigma_i X(i) = \Sigma_j Y(j) = 2S$, and S=1,002.

A_1^1	T_1^1	C_1^1	G_1^1	A_2^1	T_2^1	C_2^1	G_2^1	A_3^1	T_3^1	C_3^1	G_3^1	X(1)
A_1^2	T_1^2	C_1^2	G_1^2	A_2^2	T_2^2	C_2^2	G_2^2	A_3^2	T_3^2	C_3^2	G_3^2	X(2)
Y(1)	Y(2)	Y(3)	Y(4)	Y(5)	Y(6)	Y(7)	Y(8)	Y(9)	Y(10)	Y(11)	Y(12)	

Fig. 1. An example of comparison of the two latent triplet periodicity matrices is shown. Both matrices are present as 4N-mer vectors.

A matrix M2 was constructed as expected over a set of the random matrices, having the same marginal quantities X(i) and Y(j) as M1: $M2(i,j) = \frac{1}{2S} X(i) \text{ x } Y(j)$.

The Pearson statistics, whose value distribution follows the χ^2, allows to estimate the deviation of quantities in the M1 matrix from the expected ones in M2 matrix:

$$U = \chi^2 = \Sigma_{i,j}\{(M1(i,j) - M2(i,j))^2\}/M2(i,j).$$

The number of the χ^2 freedom degrees was 2 x 4N – 1, that is the number of comparison ranks (the number of matrix M1 or M2 elements) minus the number of independent linkages – a single claim on constancy of marginal elements: X(1) = X(2) = S.

A comparison of the original periodicity matrices was done taking into consideration all the cyclic permutations of their columns, necessary because of the uncertainty of the period start position. These permutations were adequately reflected in the original vector representations of the matrices. A possibility of classic DNA inversions was considered. In such a case, the original vector was replaced for the complementary and inverse variant. A general comparative scheme between the vectors is as follows. The first vector from a set was compared with the others, as described above, taking into consideration all the cyclic permutations, and possible inversion. The least value of the χ^2, found over all the comparisons, was fixed. If the value corresponded to chance probability of not more than 5 %, the two corresponding vectors were combined via recapitulation of their elements. Such a new vector was normalised again at the medium length of the periodic region. A cyclic permutation and fixed inverse and complementary transformation was considered in vector combination. The process of comparison between the vectors was continued until the χ^2 values corresponding to 5 % level were found. Thus, the classes of compared vectors (the periodicity matrices) was revealed.

The critical level of the χ^2 value was estimated in the result of all 2N trials in searching for pair vector similarity. A chance probability of similarity found in 2N trials: $\alpha = \mathbf{1 - (1\text{-}p)^{2N}}$ should be not more than 5 %. From this point, a critical level of chance probability in one trial **p** was counted using the inverse χ^2 function.

A selection of the unique triplet periodicity classes. We have built the relation database GECOK (Frenkel *et al.*, 2004) under DBMS Firebird to analyse the triplet periodicity of genes and the periodicity classes. Firstl, all the crossed regions of triplet periodicity in KEGG-25, which were found in the same gene, were revised. If a crossed area of periodicity regions was more than 30 % of the smallest region, region of minor statistical significance (Z-score) has been excluded from the whole list of triplet periodicity. Further, the number of cases of the same class periodicity in a single gene was equalled to the one. Finally, in the 321.085 genes from KEGG-25, where triplet periodicity was revealed, 340.079 cases of unique periodicity were determined. The unique triplet periodicities were grouped into 30.310 classes.

A distribution of quantitative content of the unique periodicity classes. Determination of a significant class size. The outlined 30.310 classes of unique triplet periodicity show a great variety in their sizes that is in the numbers of the unique periodicity matrices combined in a single class. For example, the six classes had a size ranging from the 3.000 to 4.000 matrices, and the 21.462 classes have been present by a single matrix. Considering a distribution of the original unique matrices over the 30.310 classes occur with equal probabilities, we have determined the class size can be considered as not casual with likelihood of more than 95 %. To count the critical class size we used an integral function BINOMDIST from the Microsoft Excel editor, which implement the Bernoulli scheme of trials. Under given conditions, we found the critical class size of 17 matrices of unique periodicity. The whole number of classes,of smallest than the critical size was 1169.

A comparison of the genomes' encoder "backbones" from various organisms. The outlining of not casual classes of triplet periodicity allowed us to consider the content of such a periodicity for all the 112 organisms in the KEGG database. Genome encoder "backbone" of every organism was described by distribution of its unique periodicities over the outlined 1169 classes. Thus, the total sum of the elements in such a "backbone" was always lenity. In a certain sense, every organism corresponds to a vector in the 1169-dimension of space. The distance between each pair of the organisms was determined as the square root of the sum of squared differences of the corresponding vector coordinates in the considered space. Thus, a matrix of pair distances between the organisms was obtained, reflecting the similarities of the genome encoder "backbones" among the organisms. The NEIGHBOR program, implementing the Neighbour Joining Method (Saitou, Nei, 1987), from the PHYLIP package was used for inferring the phylogeny (http://evolution.genetics.washington.edu/phylip.html).

Results and Discussion

Determination of the 1169 not casual classes of triplet periodicity in the 321,085 genes of all investigated organisms allows proposition of a basic set of the DNA sequences, whose duplications gave rise to all the known genes. These classes obviously support the suggestion that duplications following DNA divergency are the main machinery in protein evolution (Ohno, 1970). The inferred phylogeny of considered organisms on the basis of their genome encoder "backbones" similarity is shown in Figure 2.

It appears that organisms belonging to the major domains of life: Archaea (A), Bacteria (B), Eucarya (E) – form the uniform clusters. Figure 2 shows an overview omitting the details of clustering inside domain groups. Cluster 1E includes the higher eukaryotic organisms: *C. elegance*, *D. melanogaster*, *D. rerio*, *M. musculus*, *R. norvegicus*,

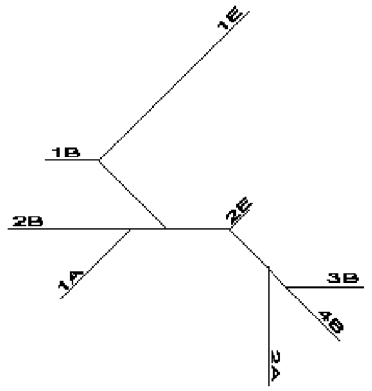


Fig. 2. A phylogenetic tree inferred by the Saitou and Nei (1987) method for the genomes' encoder "backbones" of 112 organisms. Only a general order of clustering of Archaea (A), Bacteria (B), Eucarya (E) is shown. Cluster 1E is formed by higher eukaryotes (human, plant, fly and others); Cluster 2E includes fungi (*S.cerevisiae*, *S.pombe*, *and E.cuniculi*).

H. sapience, *A. thaliana* and *P. falciparum*. Cluster 2E combines the lower Eukaryotes presented by fungic organisms. The clusters' scheme of organisms in Fig. 2 suggests the existence of specific triplet periodicity classes in each domain of life. Transfer from one domain of life to another is accompanied by exclusion of one basis and acquisition of other triplet periodicity classes. Nevertheless, a set of periodicity classes which are common among the all domains also exists.

In general, the results of our work demonstrate the fundamental significance of the triplet periodicity classes in formation of genome structure, and provide a basis for further analysis of how the obtained classes are related with the functional variety of proteins, encoded by the genes with triplet periodicity of known classes.

References

Frenkel F.E., Chaley M.B., Korotkov E.V., Skryabin K.G. Informational aspects of the latent triplet periodicity analysis // Proc. BGRS'2004. Novosibirsk.

Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method for analysis of symbolical sequences // Physical Letters A. 2003. V. 312. P. 198–210.

Ohno S. Evolution by gene duplication. Springer-Verlag, Berlin, 1970.

Saitou N., Nei M. The Neighbor-joining method: a new method for reconstructing phylogenetic trees // Mol. Biol. Evol. 1987. V. 4. P. 406–425.

MCMC METHOD HAS FOUND THAT MULTIPLE SCLEROSIS IS ASSOCIATED WITH TWO-THREE GENES COMBINATIONS

Favorov A.V.[*1], Favorova O.O.*[*2], Andreewski T.V.[2], Sudomoina M.A.[2], Alekseenkov A.D.[2], Kulakova O.G.[2], Boiko A.N.[2], Gusev E.I.[2], Parmigiani G.[3], Ochs M.F.[4]*

[1] State Scientific Centre "GosNIIGenetika", 1st Dorozhny pr., Moscow, 117545, Russia; [2] Russian State Medical University, Ostrovitianova ul., 1, Moscow, 117997, Russia; [3] Johns Hopkins University, 550 North Broadway, s. 1103, Baltimore, MD, 21205, USA; [4] Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA, 19111, USA
* Corresponding authors: e-mail: olga_favorova@mail.ru; favorov@sensi.org

Keywords: *Markov chain Monte Carlo, multiple sclerosis, polygenic disease, genetic heterogeneity, patterns extraction, allelic patterns, MCMC, MS*

Resume

Motivation: Multiple sclerosis (MS) is a typical polygenic disease. Dissection of its genetic background is complex due to the multiplicity of contributing genes, etiological and pathological heterogeneity.

Results: In a case-control study, multiple candidate genes involved in the autoimmune response were examined in common groups of Russian MS patients and unrelated controls. In the data set, a new Bayesian Markov chain Monte Carlo method identified previously found MS-associated alleles *DRB1**15(2), TNFa*9 and a biallelic combination *CCR5*Δ32,*DRB1**04, as well as two new MS-associated tri-allelic combinations: 509*TGF*β*1**C,*DRB1**18(3),*CTLA4**G and *TNF**B1,*TNF**A2,*CTLA4**G. These results are the first to show the interplay of more then two loci in conferring susceptibility to MS and its genetic heterogeneity.

Availability: The executable files for Win32 and FreeBSD implementing the method used, together with documentation, are available free to academic users upon a request.

Introduction

The genetic dissection of polygenic human diseases remains a great challenge to researchers (Threadgill *et al.*, 2002). Susceptibility to such diseases is thought to arise from the cumulative contribution of multiple independently acting and/or interacting polymorphic genes, each exerting a small or moderate effect on the overall risk. Further, a common feature of many polygenic diseases is clinical and pathological heterogeneity. Multiple sclerosis (MS) is a chronic, immune-mediated neurological disease affecting young adults, and can be considered as a prototype for polygenic human diseases (Bomprezzi *et al.*, 2003). Candidate genes for MS predisposition studies have been selected mainly because their encoded proteins are involved in autoimmune pathogenesis. To date, the HLA class II *DRB1**1501/*DQA1**0102/*DQB1**0602 haplotype is the only region repeatedly confirmed as being associated with MS in most Caucasians (Herrera, Ebers, 2003).

The present study is based on the idea that polygenic diseases are best investigated by simultaneously examining multiple candidate genes in common groups of patients and controls. Such studies may be used to search for a statistical correlation between a disease and not only distinct alleles of candidate genes but also a combination of alleles. To examine the possibility that the interplay of several genes creates risk factors for a polygenic disease, it is necessary to explore a massive number of potential combinations of allelic variants. We describe here two tri-allelic genetic combinations (patterns) reliably associated with MS in Russians that have been identified in a case-control study by a new variant of Markov chain Monte Carlo sampling.

Methods and Algorithms

286 unrelated patients with clinically defined MS, and 362 healthy unrelated controls, all of Russian descent, were genotyped at polymorphic loci at or near genes of the autoimmune inflammatory response, which are presented in the table below.

Gene/marker	Polymorphism type	Method of analysis
DRB1	Allele groups corresponding to serological specificities DR1 - DR18(3)	PCR-SSP
Microsatellite TNFa	$(AC)_n$	Nested PCR
Microsatellite TNFb	$(TC)_n$	Nested PCR
TNF	SNP –376A→G	PCR-SSP
	SNP –308G→A	PCR-SSP
	SNP –238A→G	PCR-RFLP (*Bam*HI)
LT	SNP +252G→A	PCR-RFLP (*Nco*I)
	SNP +319C→G	PCR-RFLP (*Asp*HI)
TGFβ1	SNP –509C→T	PCR-SSO
	SNP +72 wild type→C insertion	
	SNP +869T→C (10Leu→Pro)	
	SNP +915G→C (25Arg→Pro)	
	SNP +1632C→T (263Thr→Ile)	
CCR5	Wild type or a 32 base pair deletion	PCR
CTLA4	SNP +49A→G (17Thr→Ala)	PCR-RFLP (*Bst*EII)

The genotypes of all patients and controls were entered in a database, together with personal data for both controls and patients. To test correlation between the disease and an allelic pattern, original scripts performing standard statistical analysis were used.

To search for variants of alleles at multiple loci combinations that are statistically correlated with a disease, we created a new algorithm (Favorov, Ochs, 2002), allowing the Markov chain Monte Carlo (MCMC) exploration (Besag *et al.*, 1996; Robert, Casella, 2002) of genotypes tied to phenotypic trait levels. The algorithm operates by creating potential combinations (patterns) of alleles of different loci. For each locus, a pattern identifies a genotype, or an allele carriership, or lack of effect (i.e., no specific allele assigned). The patterns combine into a pattern set, with each pattern effect evaluated by a Wilcoxon-Mann-Whitney rank test in isolation (as in a statistical adjustment) from others. With each step, the algorithm proposes a new pattern from a distribution that prefers *a priori* the proposition that any locus or loci combination has no effect on the disease. For a relevant locus, the algorithm uses a Dirichlet prior distribution for the allele frequencies. Steps of the Markov chain, which form the set of Monte Carlo trials, are determined in a standard Metropolis-Hastings way (Robert, Casella, 2002) using the result of the rank test in a Bayesian framework. After an equilibration period, the Markov chain samples the sets of patterns, identifying genotypic patterns that have a high probability of being associated with disease.

Results

The MCMC algorithm identified five patterns that had a high probability of being associated with MS. All the identified patterns deal with carriership of alleles, without distinguishing homozygotes and heterozygotes. Two previously identified alleles, *DRB1**15(2) (Boiko *et al*., 2002) and TNFa*9 (Gusev *et al*., 1997) and one previously identified "duet" of alleles (*CCR5*Δ32,*DRB1**04) (Favorova *et al*., 2002) were found as reliably associated with MS. Importantly, two new patterns comprising "trios" of allelic variants were identified as reliably MS-associated as well. The first pattern contains the C allele of SNP –509 of the transforming growth factor β1 (TGFβ1) gene, *DRB1**18(3), and the G allele of the cytotoxic T-lymphocyte antigen 4 (CTLA4) gene (trio 1). The second pattern is the combination of two alleles of tumor necrosis factor (TNF) genes: –238*TNF**B1, –308*TNF**A2 and, again, *CTLA4**G (trio 2). Use of Fisher's exact tests to evaluate these findings confirmed the association of MS with the first ($p < 0.01$, OR = 17.0) and the second ($p < 0.01$, OR = 18.0) patterns.

It is important to confirm that the algorithm correctly identified each pattern as a minimal allelic set, which provides more reliable MS association than any subset of the pattern. Indeed, the phenotypic frequencies of all alleles and of all two-element subsets involved in the "trios" compared by Fisher's exact test do not differ significantly between MS patients and control individuals, confirming the necessity of all three loci to form the observed susceptibility patterns. The fact that two alleles of *TNF* gene are required to form trio 2 is in agreement with the data suggesting that *TNF**B1 and *TNF**A2 are not part of a single haplotype. Thus, the idea that genetic susceptibility to MS arises as a result of the interplay of several polymorphic genes involved in the autoimmune inflammatory response is supported.

Based on the presented results, several (partially overlapping) subgroups may be identified in the common group of Russian MS patients depending on carriership of distinct susceptibility patterns including one, two and three alleles of candidate genes. These results are evidence of the genetic heterogeneity of MS.

Discussion

Due to the linkage disequilibrium of chromosomal loci, a genetic epidemiological approach used here cannot prove unambiguously that a disease-associated gene is a causal gene of that disease. However, a biological role for the *DRB1, CCR5, TGF*β*1, TNF* and *CTLA4* gene products in the pathogenesis of MS is plausible, and supports the idea that these genes are true MS susceptibility genes.

The patterns of trios 1 and 2 have striking similarities, which may determine their MS-predisposing properties due to dysregulation of inflammatory pathways. First, both trios include the allele G of the gene for co-stimulatory molecule CTLA4, which is an important inhibitor of T-cell activation. Second, both trios include the alleles of cytokine genes which promote immune response owing to decreased level of antiinflammatory cytokine TGFβ1 (trio 1) or increased level of proinflammatory cytokine TNF (trio 2). As a whole, in trio 2 carriers, negative regulation is hit once, and signalling activity is hit twice. In trio 1 carriers, negative regulation is hit twice, coupled with an additional contribution of the HLA class II *DRB1**18(3) which is not associated alone with MS. The nature of the interplay of the alleles in trios 1 and 2 remains unclear. It may arise as a result of a threshold effect of multiple hits. However, this proposal does not exclude a possibility of unknown epistatic interactions of genes involved into the interplay of alleles belonging to each trio.

The results presented here provide the first identification of patterns requiring more than two variant alleles for a genetic predisposition to the MS as a complex, polygenic disease. The method used to identify these patterns is highly efficient, requiring only a few hours of computation time on a laptop computer for this data set, and extremely flexible, requiring only routine encoding of known genotyping data for patients and for control individuals into a symbolic code. The algorithm will therefore provide a valuable resource for the growing volume of polygenic disease-related genome data, allowing such data to be efficiently explored to identify genetic predisposition and potential therapeutic targets.

Acknowledgements

The work is partially supported by the Russian Foundation for Basic Research (grant #02-04-50041), by the Russian Federation Ministry of Industry and Science under contract #43.073.1.1.2514, by NIH (NCI CA06927), by the Pennsylvania Department of Health, and by the Pew Foundation.

References

Besag J., Green P., Higdon D., Mengersen K. Bayesian computation and stochastic systems // Statistical Science. 1996. V. 10(1). P. 3–66.

Boiko A.N., Gusev E.I., Sudomoina M.A., Alekseenkov A.D., Kulakova O.G., Bikova O.V., Maslova O.I., Guseva M.R., Boiko S. Y., Guseva M.E., Favorova O.O. Association and linkage of juvenile MS with HLA-DR2(15) in Russians // Neurology. 2002. V. 58. P. 658–660.

Bomprezzi R., Kovanen P.E., Martin R. New approaches to investigating heterogeneity in complex traits // Med. Genet. 2003. V. 40. P. 553–559.

Favorov A.V., Ochs M.F. MCMC method for identification of allelic patterns In data with quantitatively describable phenotypic features // Proceedings BGRS 2002. V. 2. P. 47–50.

Favorova O.O., Andreewski T.V., Boiko A.N., Sudomoina M.A, Alekseenkov A.D., Kulakova O.G., Slanova A.V., Gusev E.I. The chemokine receptor CCR5 deletion mutation is associated with MS in HLA-DR4-positive Russians // Neurology. 2002. V. 59. P. 1652–1655.

Grainger D.J., Heathcote K., Chiano M., Snieder H., Kemp P.R., Metcalfe J.C., Carter N.D., Spector T.D. Genetic control of the circulating concentration of transforming growth factor type beta1 // Hum. Mol. Genet. 1999. V. 8. P. 93–97.

Gusev E.I., Sudomoina Ì.À., Boiko A.N., Turetskaya R.L., Deomina T.L., Alekseev L.P,. Boldyreva M.N., Trophimov D.Yu., Undritzov I.M., Favorova O.O. TNF gene polymorphisms: association with multiple sclerosis susceptibility and severity / Eds. O. Abramsky, H. Ovadia. Frontiers in Multiple Sclerosis, London, Martin Dunitz, 1997. P. 35–41.

Herrera B.M., Ebers G.C. Progress in deciphering the genetics of multiple sclerosis // Curr. Opin. Neurol. 2003. V. 16. P. 253–258.

Robert C.P., Casella G. Monte Carlo Statistical Methods. Springer-Verlag, 2002.

Threadgill D.W., Hunter K.W., Williams R.W. Genetic dissection of complex and quantitative traits: from fantasy to reality via a community effort // Mamm. Genome. 2002. V. 13. P. 175–178.

EVOLUTION OF BACTERIAL REGULATORY SYSTEMS

Gelfand M.S.

Institute for Information Transmission Problems, RAS; State Scientific Center GosNIIGenetika; Department of Bioengineering and Bioinformatics, MSU. Moscow, Russia, e-mail: gelfand@iitp.ru

Keywords: *bacteria, regulatory systems, evolution*

Motivation: Availability of hundreds of bacterial genomes, and tens of genomes within several well-studied taxonomic groups (alpha-, beta-, and gamma-proteobacteria; Gram-positive bacteria from the Bacillus/Clostridium group), allows one to perform comparative analysis of multiple interacting regulatory systems, and, moreover, study the evolution of these systems.

Results: Comparative studies of regulation resulted in identification of riboswitches, regulatory systems based on formation of alternative RNA structures stabilized by direct binding of small molecules (vitamins, amino acids and their derivatives). Riboswitches, some of which has been subsequently studied in experiment, are direct regulators not requiring and secondary intermediaries, regulate all major cellular processes, transcription, translation and splicing, and are distributed in all three domains of life, bacteria, archaea and eukaryotes. Thus they are good candidates to the role of regulators in the RNA world and seem to be the oldest regulatory system. On the other hand, comparative studies demonstrated that transcription factor BirA regulates biotin biosynthesis in bacteria and archaea, although in this case it is not clear whether it is an ancient system predating the divergence of these domains or a result of multiple horizontal transfer events.

The evolution of regulatory systems can be studied on several levels. One can analyze the evolution of individual protein-binding sites regulating expression of orthologous genes in closely related bacteria. A surprising observation is that non-consensus nucleotides in orthologous sites from different genomes tend to be conserved.

Analysis of recognition signals in protein families (BirA, LacI, FUR and others) demonstrated that, in addition to positional changes in the consensus patterns, a frequent mode of evolution seems to be the change in the spacing between half-signals recognized by subunits of dimeric transcription factors, probably reflecting the changing geometry of the quarternary structure. Indeed, the specificity-determining positions in such factors are concentrated not only on the substrate-binding and DNA-binding surfaces, but also in the zone of contact between the subunits.

Composition of regulons also is subject to evolutionary change. The regulon may experience rapid expansion, like the FruR regulon of enteric bacteria. If a set of functionally related genes is regulated by several transcription factors, one often observes that genes regulating by one factor in some taxonomic group, are regulated by another factor in another group. Many such examples were observed in the respiration switches of gamma proteobacteria regulated by FNR (aerobic/anaerobic switch), ArcAB (sensing the oxygen awailability) and NarPL/NarQX (nitrate/nitrite respiration).

Finally, one can observe total changes in the regulation mechanism. Thus, the S-box riboswitch regulating methionine biosynthesis in the ancestor of Firmicutes was retained by bacilli and clostridia, but substituted by the expanded Met-T-box regulon in lactobacilli and by transcriptional regulator MtaR in streptococci and lactococci.

Although at present these observations are scattered and we have no general theory of the evolution of regulation, accumulation of data and development of new methods of analysis provide a variety of examples that will serve as a foundation for this theory. On the other hand, analysis of regulation is an important tool for functional annotation of proteins and metabolic reconstruction. In particular, this allows for assigning specificity for transporters and filling gaps in metabolic reconstruction due to non-orthologous gene displacement.

One recent example of such analysis was assigning a function to uncharacterized paralogs of ribosomal proteins. It was demonstrated that these proteins are regulated by the concentration of zinc and substitute the main zinc-containing proteins in the ribosome in conditions of zinc starvation. This prediction was subsequently verified in experiment.

Acknowledgements

This overview is based on results from collaborations with Anna Gerasimova, Olga Kalinina, Alexei Kazakov, Ekaterina Kotelnikova, Galina Kovaleva, Olga Laikova, Elena Lyubetskaya, Vassily Lyubetsky, Vsevolod Makeev, Andrei Mironov, Ekaterina Panina, Elizaveta Permina, Alexandra Rakhmaninova, Dmitry Ravcheev, Dmitry Rodionov, Elena Stavrovskaya, Alexei Vitreschak supported by grants from the Howard Hughes Medical Institute, the Ludwig Institute for Cancer Research, the Russian Fund of Basic Research, the Fund for Support of Russian Science, and the Program in Molecular and Cellular Biology of RAS.

References

Gelfand M.S., Laikova O.N. Prolegomena to the evolution of transcriptional regulation in bacterial genomes. Functional Genomics Series // Frontiers in Computational Genomics / Eds. M.Y.Galperin, E.V. Koonin. 2003. V. 3. P. 195–216 (Caister Academic Press).

Kalinina O.V., Mironov A.A., Gelfand M.S., Rakhmaninova A.B. Automated selection of positions determining functional specificity of proteins by comparative analysis of orthologous groups in protein families // Protein Science. 2004. V. 13. P. 443–456.

Panina E.M., Mironov A.A., Gelfand M.S. Comparative genomics of bacterial zinc regulons: Enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins // Proc. Natl Acad. Sci. USA. 2003. V. 100. P. 9912–9917.

Panina E.M., Vitreschak A.G., Mironov A.A., Gelfand M.S. Regulation of biosynthesis and transport of aromatic amino acid in low-GC Gram-positive bacteria // FEMS Microbiol. Lett. 2003. V. 222. P. 211–220.

Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes // J. Biol. Chem. 2003. V. 278. P. 41148–41159.

Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. Regulation of lysine biosynthesis and transport in bacteria: yet another riboswitch // Nucleic Acids Res. 2003. V. 31. P. 6748–6757.

Rodionov D.A., Vitreschak A.G., Mironov A.A., Gelfand M.S. Comparative genomics of the regulation of methionine metabolism in Gram-positive bacteria // Nucleic Acids Res. 2004. (in press).

Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element // RNA. 2003. V. 9. P. 1084–1097.

Vitreschak A.G., Rodionov D.A., Mironov A.A., Gelfand M.S. Riboswitches: the oldest mechanism for the regulation of gene expression? // Trends in Genetics. 2004. V. 20. P. 44–50.

Vitreschak A.G., Lyubetskaya E.V., Shirshin M.A., Gelfand M.S., Lyubetsky V.A. Attenuation regulation of amino acid biosynthetic operons in proteobacteria: comparative genomics analysis // FEMS Microbiol. Lett. 2004. (in press).

Zhang Z., Feige J.N., Chang A.B., Anderson I.J., Brodianski V.M., Vitreschak A.G., Gelfand M.S., Saier M.H. Jr. A transporter of Escherichia coli specific for L- and D-methionine is the prototype for a new family within the ABC superfamily // Arch. Microbiol. 2003. V. 180. P. 88–100.

COMPARATIVE GENOMIC ANALYSIS OF RESPIRATION SWITCH IN GAMMA-PROTEOBACTERIA

Gerasimova A.V.[1], *Ravcheyev D.A.**[2, 3], *Gelfand M.S.*[1, 2, 3], *Rakhmaninova A.B.*[2]

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

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

Keywords: *comparative genomics, gamma-proteobacteria, aerobic and anaerobic respiration, FNR, ArcA, NarL, NarP*

Summary

Motivation: Selection of respiration pathways in prokaryotes appears to be one of the most important aspects of energy metabolism. Bacteria have a great number of respiratory enzymes and demonstrate then complicated regulatory pattern. Thus, it is of in interest to analyze the regulation of respiration in different species. Here, we studied the group of gamma-proteobacteria.

Results: A large number of gamma-proteobacteria genomes was studied by the methods of comparative genomics. Putative ArcA, FNR and NarP regulons were described in detail for a number of microorganisms. After that different regulons for each organism were compared. We also confirmed known FNR-binding signal, improved the recognition profile for the ArcA-signal and created a new profile for the NarP-signal.

Introduction

Prokaryotes, for instance *Escherichia coli*, can adapt to a wide variety of environmental conditions. A source of this capability is the presence of numerous aerobic and anaerobic respiratory systems. To adapt to different growth conditions, bacteria alter the composition of their respiratory systems by changing substrate-specific dehydrogenases and terminal oxidoreductases. The concentration of each component is strictly regulated in order to optimize the respiratory chains according to the substrates present and the physiological needs of the cell.

In *E. coli*, the switch between aerobic and anaerobic metabolism is controlled by transcription factors FNR and ArcA. Both proteins may act as repressors and activators depending on the functions of regulated genes. FNR was shown to be active as a dimer undergrownd anaerobiosis. Conversely, ArcA is activated in response to oxygen availability. It is also well known that the *arcA* gene may be repressed or activated by FNR dependent at the oxygen level. Thus these two proteins can activate systems for one type of respiration and at the same time repress expression of genes essential for metabolism of another type.

Further level of regulation is provided by homologous regulators NarL and NarP in response to the availability of nitrate and nitrite under anaerobic conditions. Nitrate is the preferred electron acceptor for anaerobically growing cells because of relatively high redox potential of the nitrate/nitrite couple. This regulatory system activates genes for nitrate- and nitrite-specific terminal reductases and represses synthesis of enzymes for alternative types of anaerobic respiration. The response to nitrate and nitrite is mediated by two homologous sensors NarX and NarQ. Duplication of this system seems to be due to the need for fine tuning of the nitrate and nitrite reducing systems in response to a dynamic ratio of two alternative electron acceptors (Gennis, Stewart, 1996).

Despite the fact that physiology and regulation of respiration in *E. coli* and its close relatives was studied indepth, many crucial problems remain unsolved. Only FNR regulon was well studied, by different approaches including comparative genomics approaches including (Gerasimova *et al.*, 2001). Regulation by ArcA is poorly investigated and information about its interaction with DNA is rather scarce. The data on the NarL and NarP regulatory system are very controversial. Although this system was intensely studied during the last two decades, the mechanisms of crossregulation and the structure of the binding signals remained elusive. Moreover, complex analysis of regulation by all the four factors was done for only a few operons.

Methods

Complete genome sequences of *Escherichia coli K12*, *Salmonella typhi Ty2*, *Yersinia pestis KIM*, *Haemophilus ducreyi 35000HP*, *Haemophilus influenzae Rd KW20*, *Pasteurella multocida*, *Vibrio cholerae O1*, *Vibrio parahaemolyticus RIMD 2210633*, and *Vibrio vulnificus CMCP6* were downloaded from the GenBank. The complete sequence of *Yersinia enterocolitica* was extracted from the Sanger Institute web site (http://www.sanger.ac.uk/). Partial sequence of *Actinobacillus actinomycetemcomitans HK1651* was extracted from the University of Oklahoma's Advanced Center for Genome Technology web site (http://www.genome.ou.edu/act.html). Partial sequence of *Vibrio fischeri ES114* was extracted from the GOLD Genomes OnLine Database web site (http://ergo.integratedgenomics.com/GOLD). The comparative genomics methods were used. This approach was described in detail in Mironov *et al.* (1999).

Candidate sites were identified in the upstream regions of the annotated genes, including the predicted. A gene was considered as a member of a regulon, if it had an upstream candidate site (-400 to +50 nucleotides relative to the gene start) conserved upstream of the orthologous from related organisms. Genes were considered to belong to one operon, if they were transcribed in the same direction and the intergenic distances did not exceed 100 nucleotides.

In this work, we compared pairwise the genomes belonging to a single taxonomic group, for example, the Pasteurellaceae or Vibrionaceae families. This approach provides an opportunity for the recognition of group-specific features of regulation. For the identification of orthologous genes and regulatory recognition patterns, the Genome Explorer program (Mironov *et al.*, 2000) was used. The SignalX program (Mironov *et al.*, 2000) was used to create recognition profiles. Multiple sequence alignment and construction of phylogenetic trees were carried out by using the ClustalX program (Thompson *et al.*, 1997). Sequence logos were constructed using the MakeLogo program (Schneider, Stephens, 1990) as implemented in (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

Results and Discussion

Duplicated nitrate-responsive regulators. Comparative analysis demonstrated that duplication of this system occurred just before the divergence of Enterobacteriaceae, Pasteurellaceae and Vibrionaceae branches, but duplicated regulators and sensors were conserved only in *E. coli* and *Salmonella typhi*. In all the genomes, NarL co-occurs with cognate sensor NarX, whereas NarP co-occurs with NarQ in all the genomes except the Yersinia genus, where NarP and NarX are present. In this case, some changes in amono acid sequence of signal transduction proteins were observed.

Construction of recognition profiles for the ArcA and NarP binding sites. For the analysis of the FNR regulon, we used the previously described recognition profile (Gerasimova *et al.*, 2001).

For the recognition of the ArcA-binding sites, we employed the profile generated by the SeSiMCMC program (Gerasimova *et al.*, 2004).

Previously, we demonstrated that the *narP* gene always co-occurs with operons for the periplasmic nitrate and nitrite reduction system *nap*, *nrf* and *ccm*. In the genomes containing NarP but lacking NarL, all these operons were preceded by 16-bp palindromic sites. Each half of this palindrome corresponds to the predicted consensus for the NarL/NarP-binding signal.

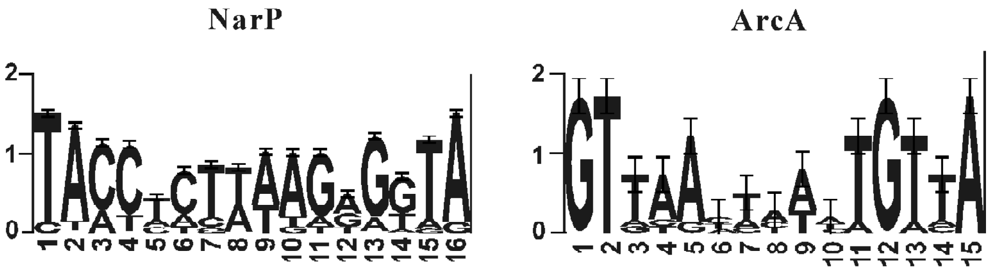


Fig. Sequence logos for the NarP and ArcA binding sites. Horizontal axis, position in the binding site; vertical axis, information content in bits. The height of each stack of letters is proportional to the positional information content at the given position; the height of each individual letter reflects its prevalence at the given position.

Complex analysis of regulons. The genomes of organisms from three gamma-proteobacteria families: Enterobacteriaceae (*Y. pestis, Y.entercolitica*), Pasteurellaceae (*P. multocida, A. actinomycetemcomitans, H. influenzae*, *H. ducreyi*) and Vibrionaceae (*V. cholerae, V. parahaemolyticus, V. vulnificus, V. fischeri*), were investigated.

The *fnr* gene was found to be autoregulated in all the studied genomes. In Vibrionaceae and Pasteurellaceae, the putative ArcA-binding sites were also found upstream of the *arcA* genes. Moreover, in *P. multocida, A. actinomycetemcomitans*, and *H. influenzae*, the *fnr* gene was preceded by an additional NarP-site. Thus, the *E. coli* hierarchy of regulators is not conserved in the other families of gamma-proteobacteria.

In all vibrio, the *narQ* and *narP* genes form potential operon preceded by candidate sites for all the three regulators. No conserved putative sites were found upstream of single *narQ* genes whereas sites for at least one regulator were identified upstream of the *narP* genes in all the genomes.

In all genomes, NarP regulates all the orthologs of the *E. coli* NarL/NarP regulon members with the exception of some marginal regulon members. Sites upstream of these genes were found in a few genomes. The core of the NarP regulon consists of the operons for nitrate (*napFDAGHBC*) and nitrite (*nrfABCDEFG* and *nirBDC-cysG*) reductases as well as the *ccm* operon implicated in the heme export. All of these operons, except the *ccm* operon in *Y. pestis*, preceded by candidate FNR- and sometimes by additional ArcA-binding sites.

New members of the considered regulons were identified. In most cases, putative sites for all the three regulators were found upstream of the *cydAB* operon. The product of this operon is a terminal reductase included in the aerobic respiratory chain. The importance of this enzyme to the global cell metabolism explains a large number of regulatory sites upstream of the *cydAB* operon.

In all the studied genomes, the *nqr* operon has candidate binding sites of at least one of regulators. The product of this operon is a NADH-dehydrogenase. It is unrelated to the product of the *E. coli* FNR-, ArcA- and NarL/NarP-regulated operon *nuo* performing the same function.

A similar situation was observed for the molybdopterin cofactor synthesis genes *moaABCDE*. The molibdenium cofactor is an essential part of a large number of prokaryotic oxidoreductases. This explains why candidate sites for at least one of the three regulators were found upstream of the *moa* operons in all the genomes.

A more detailed description of these regulons, as well as investigation of the NarL and TorR (nitric oxide derivatives dependent respiration) regulatory systems, is the subject of further research.

Acknowledgements

We are grateful to Alexander Favorov and Andrey Mironov for kindly provided software. This study was supported by grants from the Howard Hughes Medical Institute (55000309), the Fund for Support of the Russian Science and the program in Molecular and Cellular Biology of the Russian Academy of Science.

References

Gelfand M.S., Koonin E.V., Mironov A.A. Prediction of transcription regulatory sites in Archaea by a comparative genomic approach // Nucleic Acids Res. 2000. V. 28. P. 695–705.

Gennis R.B., Stewart V. Respiration // *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology / Eds. Neidhart F.C. ASM Press, Washington, 1996. P. 217–261.

Gerasimova A.V., Rodionov D.A., Mironov A.A, Gelfand M.S. Computer analysis of regulatory signals in bacterial genomes. Fnr binding segments // Mol. Biol. 2001. V. 35. P. 1001–1009.

Gerasimova A.V., Gelfand M.S., Makeev V.U., Mironov A.A., Favorov A.V. ArcA regulator of gamma-proteobacteria: identification of the binding signal and description of the regulon // Biophysics. 2004, in press.

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

Schneider T.D., Stephens R.M. Sequence logos: a new way to display consensus sequences // Nucleic Acids Res. 1990. V. 18. P. 6097–6100.

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

GENOME-WIDE IDENTIFICATION OF MITOCHONDRIAL DNA TOPOISOMERASE I IN ARABIDOPSIS

Katyshev A.I.*[1], *Rogozin I.B.*[2], *Konstantinov Y.M.*[1]

[1] Siberian Institute of Plant Physiology and Biochemistry SB RAS, Irkutsk, Russia; [2] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Corresponding author: e-mail: yukon@sifibr.irk.ru

Keywords: *DNA topoisomerase I, mitochondria, chloroplasts*

Summary

Topoisomerases are conserved enzymes that play an important role in multiple cellular processes such as DNA recombination, DNA replication, and cell cycle checkpoint control (Pommier, 1998). It is likely that at least three different plant topoisomerases function within nucleus, in mitochondria and in chloroplasts. This hypothesis has been partially supported by biochemical experiments (Daniell *et al*., 1995; Balestrazzi *et al*., 2000). Moreover, in chloroplasts or mitochondria may function enzymes which are of different genetic origin and structure, prokaryotic topo IA and eukaryotic topo IB type topoisomerases. Thus, there may be more than one gene encoding chloroplast and mitochondrial topo I in plants. Genome-wide analysis of the *Arabidopsis thaliana* genome suggested that there is only one gene encoding non-nuclear topo I. Some explanations of this fact are provided and unique features of the gene product are discussed.

Introduction

Members of the topoisomerase I superfamily are well characterized in prokaryotes and animals. In plant nuclei such mechanisms are well known and they are similar to those in animals, but it is not a case of mitochondria and chloroplasts. Moreover, the coexistence of three genomes may affect the functioning of topo I enzymes. In our laboratory we are investigating various enzymatic and non-enzymatic factors implicated in genetic processes in plant mitochondria. Previously we showed that maize nuclear and mitochondrial topo I DNA-binding and relaxation activities are regulated in different ways (Konstantinov *et al*., 2003). This result suggested that the enzymes have different structures. To verify this hypothesis we carried out database searches of topo I homologs in the *Arabidopsis thaliana* genome. Three candidate genes were identified. Two of them are highly similar to each other and encode a nuclear topo I. The third gene encode a candidate mitochondrial / chloroplast topo I. It is likely that this gene encodes mitochondrial enzyme, however we cannot reject a hypothesis that this gene encodes a chloroplast topo I. The possible theoretical explanations of the existence of only one gene for organellar topo I are: a) enzyme contains a dual-targeting sequence allowing it to be transported into both compartments; b) topo I functions in chloroplasts are completely substituted by topo III functions; c) chloroplast topo I sequence does not have a detectable similarity to known topo I. The most plausible are the first two considerations, the data proving their validity will be discussed.

The candidate mitochondrial / chloroplast topo I is a prokaryotic type IA topoisomerase, which is also known as a *Bacillus subtilis*-like DNA topoisomerase I or eukaryotic toposimerase IIIa (KOG1956, http://ncbi.nlm.nih.gov/COG/new/).

Methods

Protein sequences of Arabidopsis topo I and topo I-like enzymes were obtained from the TAIR (The Arabidopsis Information Resource) peptide dataset (ftp://tairpub:tairpub@ftp.arabidopsis.org/home/tair/Sequences/blast_datasets/ATH1.pep.01072002). Sequences of topo I and III from other organisms used in alignment were obtained from the GenBank database (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=Search&DB=protein) and the COG/KOG database (http://ncbi.nlm.nih.gov/COG/new/). COGs/KOGs (Clusters of Orthologous Groups of prokaryotic and eukaryotic proteins) were constructed from the results of all-against-all BLAST comparison of proteins encoded in complete genomes by detecting consistent sets of genome-specific best hits (Koonin *et al.*, 2004). Alignments of topo I proteins were constructed using a ClustalW algorithm-based program from the Vector NTI5 package (Bethesda Inc., USA). Neighbor-joining trees were constructed using the MEGA2 program (http://www.megasoftware.net) (Kumar *et al.*, 2001).

To predict subcellular localization of proteins of interest we used Internet resources available at the http://www.expasy.org/ molecular biology tools server: a) the MITOPROT program (Claros, Vincens, 1996) at the http://www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter*;* b) the Predotar program at the http://www.inra.fr/predotar/; c) the TargetP V1.0 program (Emanuelsson *et al.*, 2000) at the http://www.cbs.dtu.dk/services/TargetP/; d) the PSORT program (Nakai, Kanehisa, 1991) at the http://psort.nibb.ac.jp/form.html. Data on domen organization of proteins and their motif structure were obtained by InterProScan sequence search package (http://www.ebi.ac.uk/InterProScan/) and SMART research tool (http://smart.embl-heidelberg.de/).

Results and Discussion

The Arabidopsis genome contains two highly homologous genes encoding nuclear topo I, located near each other on chromosome 5 (GenBank accession no. NP_200342, P30181). The significance of topo I enzymes in cell genetic processes is well known, and the existence of duplicated genes for nuclear topo I in Arabidopsis supports importance of this enzyme. Hence, it is reasonable to expect additional genes encoding chloroplast and mitochondrial enzymes. Interestingly, BLAST searches suggested only one candidate gene for mitochondrial / chloroplast enzyme in the Arabidopsis genome (GenBank accession no. NP_194849).

Alignments of these three proteins with amino acid sequences of topo I from other organisms indicated that the candidate gene for mitochondrial / chloroplast enzyme is of a prokaryotic origin whereas the nuclear enzymes are eukaryotic type IB topoisomerases. This observation suggested the organellar localization of the candidate gene for mitochondrial / chloroplast topo I which is also known as the *Bacillus subtilis*-like topoisomerase I. The term *Bacillus subtilis*-like topoisomerase I was proposed by Brandt *et al.* (unpublished). To determine localization of *Bsu*-like topo I in cells we used programs for prediction of nuclear/mitochondrial/chloroplast targeting signals in protein molecules. Results of signal searches are shown in the Table.

Table. Probabilities of mitochondrial and chloroplast localization of *Bacillus subtilis*-like topoisomerase I predicted by different programs

Name of program	Probability of mitochondrial localization	Probability of chloroplast localization
MITOPROT	0.9996	-
Predotar	0.822	0.001
TargetP V.1.0	0.759	-
PSORT	0.861	0.870

These results strongly suggest mitochondrial localization of the enzyme. However, the PSORT program suggested that the *Bsu*-like topo I is a dual-targeted enzyme which functions also in chloroplasts. There are some proteins of dual-targeting in Arabidopsis (Chow *et al.*, 1997). However, we cannot exclude that a topo I enzyme was substituted by a topo III enzyme in chloroplasts.

Phylogenetic analysis of prokaryotic topo I orthologs (COG0550, http://www.ncbi.nlm.nih.gov/COG/new/release/cow.cgi?cog=COG0550) suggested that the candidate mitochondrial topoisomerase I may be acquired from alpha-proteobacteria (Fig. 1) which are believed to be ancestors of eukaryotic mitochondria. This result is an additional indication that the *Bsu*-like Arabidopsis topo I may still function in mitochondria.

The primary structure of the *Bsu*-like topo I from Arabidopsis contains several domains characteristic for prokaryotic topo I (SM00436, SM00437, SM00493). The enzyme contains a conservative prokaryotic topo I functional motif (ProSite accession no. PS00396) adjacent to the active Tyr residue

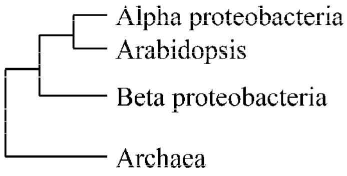


Fig. 1. Schematic representation of a phylogenetic tree for topoisomerase I homologs.

(Fig. 2) in the core domen indicating that this is functional DNA topo I enzyme.

In conclusion, genome-wide search results allowed us to identify the gene encoding the candidate mitochondrial DNA topoisomerase I in *Arabidopsis thaliana*. A chloroplast topo I gene is still not found. The experimental data should be obtained to verify hypothesis on dual-targeting of *Bsu*-like topo I.

```
S.cer.   LNAKQSLDAAEKLYQ---------KGFIS Y PRTETDTFPH-AMDL-KSLVEKQAQLDQLAAGGRTAWASY    394
E.coli   FGVKKTMMMAQRLYE---------AGYIT Y MRTDSTNLSQDAVNMVRGYISDN----------------F    343
H.inf.   FGVKKTMMLAORLYE---------AGYIT Y MRTDSTNLSODALNMARSYIENH----------------F    349
B.sub.   FRAKKTMMIAQQLYEGIDLGREGTVGLIT Y MRTDSTRISNTAVDEAAAFIDQT----------------Y    322
A.th.    FSTAHTMKLAQKLYEGVQLSDGKSAGLIT Y MRTDGLHIADEAIKDIQSLVAER----------------Y    801
```

Fig. 2. The alignment of prokaryotic Topo I protein sequence fragments, containing conservative functional motif, with the same in *Bacillus subtilis*-like topo I from *Arabidopsis thaliana*. *B. sub* – protein sequence fragment of topo I from *Bacillus subtilis*, *E. coli* – from *Escherichia coli*, H. inf – from *Haemophilus influenzae*, A.th (Bsu-like) – from *Arabidopsis thaliana*. The active Tyr residue is underlined.

Acknowledgements

The work was supported by RFBR Grant No. 01-04-48162.

References

Balestrazzi A., Chini A., Bernacchia G., Bracci A., Luccarini G., Cella R., Carbonera D. Carrot cells contain two top1 genes having the coding capacity for two distinct DNA topoisomerases // J. Exp. Bot. 2000. V. 51(353). P. 1979–90.

Chow K.-S., Singh D.P., Roper J.M., Smith A.G. A single precursor protein for ferrochelatase-I from *Arabidopsis* is imported *in vitro* into both chloroplasts and mitochondria // J Biol. Chem. 1997. V. 272. P. 27565–27571.

Claros M.G., Vincens P. Computational method to predict mitochondrially imported proteins and their targeting sequences // Eur. J. Biochem. 1996. V. 241. P. 779–786.

Daniell H., Zheng D., Nielsen B.L. Isolation and characterization of an *in vitro* DNA replication system from maize mitochondria // Biochem. Biophys. Res. Commun. 1995. V. 208(1). P. 287–94.

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

Konstantinov Y., Katyshev A., Subota I., Tarasenko V. Effects of redox conditions on DNA binding activity of mitochondrial topoisomerase I // Maize Gen. Coop. Newslett. 2003. V. 77. P. 37–38.

Koonin E.V., Fedorova N.D., Jackson J.D., Jacobs A.R., Krylov D.M., Makarova K.S., Mazumder R., Mekhedov S.L., Nikolskaya A.N., Rao B.S., Rogozin I.B., Smirnov S., Sorokin A.V., Sverdlov A.V., Vasudevan S., Wolf Y.I., Yin J.J., Natale D.A. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes // Genome Biol. 2004. V. 5. R7.1–R7.28 (genomebiology.com/2004/5/2/R7).

Kumar S., Tamura K., Jakobsen I.B., Nei M. MEGA2: molecular evolutionary genetics analysis software // Bioinformatics. 2001. V. 17. P. 1244–1245.

Nakai K., Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria // Proteins. 1991. V. 11. P. 95–110.

Pommier Y. Diversity of DNA topoisomerases I and inhibitors // Biochimie. 1998. V. 80. P. 255–70.

In honour of Emil Ginsburg

A SIMPLE TEST FOR LINKAGE DISEQUILIBRIUM BETWEEN A MARKER ALLELE AND A GENE MUTATION IN HETEROZYGOTE CARRIERS

Korostishevsky M.*, Bonne-Tamir B.

Dept. of Human Genetics and Molecular medicine, Sackler School of Medicine, Tel-Aviv University, Ramat-Aviv 69978, Israel

* Corresponding author: e-mail: korost@post.tau.ac.il

Keywords: *linkage disequilibrium, statistical test, mutation origin, microevolution*

Summary

Motivation: Screening for gene mutation carriers is an essential part of genetic studies. If a founder effect is assumed for a mutation, linkage disequilibrium between the mutation and adjacent markers should be estimated. The difficulty is that in most cases, frequencies of marker alleles in carrier chromosomes and in "normal" chromosomes cannot be established by direct counting. To avoid this difficulty a complex statistical methods using a control sample of non-carriers has been adopted.

Results: A simple one step statistical procedure to test linkage disequilibrium in a widespread case of heterozygote carrier sample is proposed. The method does not use a control sample and does not require a preliminary estimation of the haplotype frequencies or mutation carrier frequency.

Introduction

Screening for gene mutation carriers is an essential part of genetic studies. If a founder effect is assumed for a mutation (e.g. a disease allele), linkage disequilibrium (LD) between the mutation and adjacent markers should be estimated. The difficulty is that in most cases, frequencies of marker alleles in carrier chromosomes and in "normal" chromosomes cannot be established by direct counting. To avoid this difficulty a three-step procedure using a control sample of non-carriers has been adopted (Terwilliger, Ott, 1994):

Count the allele frequencies in a control sample of non-carriers;

Estimate the allele frequencies in carrier chromosomes using a sample of carriers and the result of the first step;

Test the difference between these two frequency sets.

This method is based on a preliminary estimation of haplotype frequencies (marker allele – mutation state). Accuracy of these estimates should be taken into account in the statistical procedure of null hypothesis rejection. A control sample is reliable only if it gives marker allele frequencies which are the same as frequencies from the normal allele of the carrier sample. Therefore, such a comparison is warranted.

We show here that a complexity in linkage disequilibrium testing could be avoided at least in one specific case, a widespread event of heterozygote carrier sample (i.e. carriers of the mutation on one allele). For example, screening for carriers of known rare mutations, often results in such samples. In addition, screening for carriers of a rare dominant disease (i.e. carriers of an unknown mutation) almost certainly results in heterozygote carrier samples. We propose a simple one step statistical procedure to test linkage disequilibrium in this case. The method does not use a control sample and does not require a preliminary estimation of the haplotype frequencies or mutation carrier frequency.

Statistical model

Let N be a number of individuals in a sample of heterozygote carriers, and N_{AA}, N_{Aa}, N_{aa} are numbers of individuals with one of the three genotypes of a marker; $N = N_{AA} + N_{Aa} + N_{aa}$. The marker allele A is hypothesized to be associated with the gene mutation and a, denotes an alternative allele or alleles.

Assuming that each of the N individual chromosome pairs represents one normal and one carrier chromosome, the likelihood of the observation is proportional to:

$$L\{p,q\} \propto [pq]^{\mathrm{NAA}}[p(1-q)+(1-p)q]^{\mathrm{NAa}}[(1-p)(1-q)]^{\mathrm{Naa}}, \quad (1)$$

where p is the unknown frequency of the marker allele A on the carrier chromosomes in the population and q is the unknown frequency of the same allele on the normal chromosomes in the population. The linkage disequilibrium hypothesis is **H**: $p \neq q$ and the null hypothesis is $\mathbf{H_0}$: $p = q$.

Note that $L\{p, q\} = L\{q, p\}$. Therefore the cases $p > q$ and $p < q$ cannot be distinguished. $L\{p, q\}$ reaches maximum at $p \neq q$ if and only if the number of heterozygotes, N_{Aa}, exceeds the Hardy-Weinberg expectation, i.e. $\Delta = (N_{Aa})^2 – 4N_{AA}N_{aa} > 0$.

In the case of no excess of heterozygotes, $\Delta \leq 0$, $L\{p, q\}$ reaches the maximum at $p = q = p_N$, where p_N is the frequency of allele A in the sample, $p_N = (2N_{AA}+N_{Aa})/2N$. Therefore, if $\Delta \leq 0$, the complex **H** hypothesis has no prevalence compared to the simple $\mathbf{H_0}$ hypothesis. Then the **H** hypothesis should be rejected.

In the case of excess hetrozygotes, $\Delta > 0$, the single maximum (in the parameter area: $0 \leq p \leq 1$, $0 \leq q \leq 1$) of $L\{p,q\}$ is reached at $p = p_N + \sqrt{\Delta}/2N$ and $q = p_N - \sqrt{\Delta}/2N$. In this case the **H** and $\mathbf{H_0}$ hypotheses can be compared statistically by using the likelihood ratio test (LRT):

$$\mathrm{LRT} = -2 * ln(L\{p_N, p_N\}/ L\{p_N + \sqrt{\Delta}/2N, p_N - \sqrt{\Delta}/2N\}) \quad (2)$$

If $\mathbf{H_0}$ is true, the LRT distribution can be approximated by the χ^2 distribution at one degree of freedom (Kendall, Stuart, 1973). Accuracy of the approximation depends on the sample size and the marker heterozygosity. The exact values of the type I error (the chance to reject true $\mathbf{H_0}$ hypothesis) are demonstrated on Figure. These values were obtained as an occurrence of LRT > 3.841 among cases where $\Delta > 0$. Probability of each of the possible sample case was evaluated as

$$P\{N_{AA}, N_{Aa}, N_{aa}\} = \frac{N!}{N_{AA}!N_{Aa}!N_{aa}!}[p]^{2\mathrm{NAA}}[2p(1-p)]^{\mathrm{NAa}}[(1-p)]^{2\mathrm{Naa}}. \quad (3)$$

The result shows that for a sample size exceeding 30 individuals and for marker heterozygosity exceeding 0.4 the χ^2 approximation is acceptable.

It should be noted that the resulting LRT formula (2), designed for LD testing in heterozygote carriers, coincides with a formula used for HWE testing (Weir, 1996). The difference between the tests is that an excess of marker heterozygotes precedes the LD testing, while the p-value corresponds to the one-side statistic criteria. However, the LD and HWE tests are based on two different statistic models; their convergence on heterozygote carrier sample is demonstrated for the first time.

Application test examples

1. Thirty-four Ashkenazi Jews who were heterozygote carriers of the 167delT mutation in the connexin 26(GJB2) gene were typed for the D13S141 marker (Morell *et al.*, 1998; Sobe *et al.*, 199). A conclusion of LD was based on an additional sample of 381 non-carriers and the estimation of the haplotype (marker allele – mutation state) frequencies in the carrier sample. The allele 3 was found to be associated with 167delT mutation in Ashkenazi Jews.

The genotypes are distributed in the carrier sample as N_{AA}=12, N_{Aa}=20, N_{aa}=2 *(**A** =* allele 3). In this case, we calculate an LRT value of 3.02. Therefore, an LD can be suggested at p ≈ 0.08, based on the carrier sample only.

2. Seventy-seven Ashkenazi Jews who were heterozygote carriers of the G197del mutation in the LDLR gene were typed for three markers: D19S221, D19S865, D19S413 (Durst *et al.*, 2001). The authors indicate that: "Haplotype data were obtained either from informative pedigrees (via cosegregation of microsatellite alleles) or via homozygosity at marker loci." About half of the carriers for which haplotypes could not be reconstructed were omitted in the LD analyses. LD conclusion was based on an additional sample of 90 non-carriers. The result for each of the three markers was presented as: "LD index values show higher degree of allelic excess at D19S221 (allele 104, δ = .808, 90 %CI: .635 – .981) and D19S413 (allele 74, δ = .868, 90 %CI: .748 – .988) than at D19S865 (allele 208, δ = .618, 90 %CI: .455 – .780)".

The genotypes at D19S221 are distributed in the complete carrier sample as N_{AA}=17, N_{Aa}=50, N_{aa}=10 (***A*** = allele 104). In this case we calculate an LRT value of 7.54. LD can be suggested at p ≈ 0.006. The genotypes at D19S413 are distributed as N_{AA}=22, N_{Aa}=45, N_{aa}=10 (***A*** = allele 74). The LRT values for D19S413 equals 3.06 and LD can be suggested at p ≈ 0.08. The genotypes at D19S865 are distributed as N_{AA}=22, N_{Aa}=38, N_{aa}=16 (***A*** = allele 208). The LRT value for D19S865 equals 0.00 and thus does not demonstrate any LD in the complete carrier sample.

The discrepancy at the D19S865 marker can be caused by a disagreement between control and carrier samples. On the one hand, the authors reported the frequency of the 208 allele in the control sample as 0.222. On the other hand, 22 out of the 76 carriers have the 208/208 genotype. Hence, the frequency of the 208 allele on "normal" chromosomes in the carrier sample is no less than 22/76 = 0.289. The ML estimate of this frequency equals 0.539. Such a difference between control and carrier sample could lead to erroneous LD conclusion.

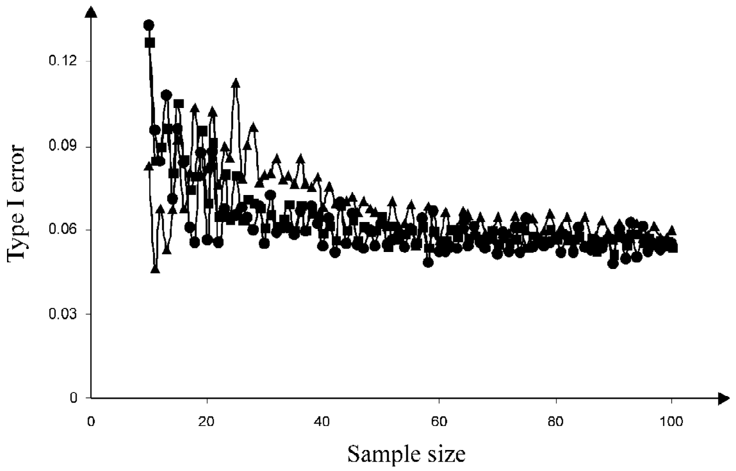


Fig. Type I error as a function of the sample size (at critical value = 3.841): three cases of different marker heterozygosity, *h* = 2*p*(1- *p*), were marked as: ▲ = 0.42, ■ = 0.48, ● = 0.50.

Acknowlegements

We thank Morell RJ., Leitersdorf E. for providing us with the marker genotypes.

References

Durst R., Colombo R., Shpitzen S., Ben Avi L., Friedlander Y., Wexler R., Raal F.J., Marias D.A., Defesche J.C., Mamdelshtam M.Y., Kotze M.J., Leitersdorf E., Meiner V. Recent origin and spread of a common Lithuanian mutation, G197del LDLR, causing familial hypercholesterolemia: positive selection is not always necessary to account for disease incidence among Ashkenazi Jews // Am. J. of Hum. Genet. 2001. V. 68. P. 1172–1188.

Kendall M.G., Stuart A. The Advanced Theory of Statistics. London: Griffin and Company Ltd, 1973. V. 2. P. 234–237.

Morell R.J., Kim H.J., Hood L.J., Goforth L.G., Friderici K., Fisher R., Camp G.V., Berlin C.I., Oddoux C., Oster H., Keats B., Friedman T.B. Mutations in the connexin 26 gene (GJB2) among Ashkenazi Jews with nonsyndromic recessive deafness // The New England J. of Medicine. 1998. V. 339. P. 1550–1555.

Sobe T., Erlich P., Korostishevsky M., Vreugde S., Shohat M., Avraham K.B., Bonne-Tamir B. High frequency of the defness associated 167delT mutation in the connexin 26 (GJB2) gene in Israeli Ashkenazim // Am. J. of Med. Genet. 1999. V. 86. P. 499–500.

Terwilliger J.D., Ott J. Handbook of Human Genetic Linkage. London: The Johns Hopkins University Press Ltd, 1994. P. 199–203.

Weir B.S. Genetic data analysis II. Sunderland: Sinauer Associates Inc, 1996. P. 101–103.

EVOLUTIONARY TREE RECONSTRUCTION AND TRAVELING SALESMAN PROBLEM: A POWERFUL ALGORITHM FOR SHAGGY TREES

***Korostishevsky M.*[1]*, *Burd A.*[2], *Mester D.*[2], *Bonne'-Tamir B.*[1], *Nevo E.*[2], *Korol A.*[2]**

[1] Department of Human Genetics and Molecular Medicine, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel 69978; [2] Institute of Evolution, University of Haifa, Haifa, Israel 31905
* Corresponding author: e-mail: korost@post.tau.ac.il

Keywords: *evolution, genetic trees, computer analysis*

Summary

Motivation: The correct Evolutionary Tree Reconstruction (ETR) based on the current genetic data is considered as an NP-complete problem. The number of "possible" trees rapidly grows with the number of "leaves". The full ETR model includes the number of ancestor vertices and the number of mutation events on the origin tree. Efficient algorithms are needed to meet the challenges of current molecular evolution studies in order to allow simultaneous treatment of hundreds and thousands of individual genotypes.

Results: We present a new ETR approach based both on the reduction the ETR problem to Traveling Salesman Problem (TSP) and on the minimization of the ETR model using Guided Evolution Strategy algorithm. The robustness of the model is defined by simulation experiments. The duration time on an ordinary computer, Pentium-4, is a few seconds for several hundreds of leaves.

Availability: http://study.haifa.ac.il/~aburd/genetic.html

Introduction

In any evolutionary process, speciation events cause a new species to split off from an existing one, thus creating the diversity of life forms we know today. A key issue in evolutionary biology is to reconstruct the history of these speciation events. An evolutionary tree is a rooted tree, where each internal vertex has at least two descendent vertices and the final vertices are labeled with distinct symbols representing recent species. The goal is: given the properties of the recent species, reconstruct what the tree is. Much of the current interest in evolutionary trees derives from the increasing availability of DNA sequence data. We consider here situations where the data represent non-recombining DNA sequences, such as DNA of Y-chromosome or mitochondrial DNA, where the origin for all sites of the given sequence is the same and the speciation events are defined only by the mutation process (Bonne-'Tamir *et al.*, 2003).

The evolutionary tree is an oriented graph. A set of DNA sequences {s} defines the vertices of a tree. Length of an edge is defined by the number of generations between vertices. Equal number of generations is assumed from the root vertex to any final vertex. To a tree which contains n vertices, we shall number vertex with natural numbers 1, ..., n. To each vertex i of the tree, we define two corresponding numbers a_i and g_i, where a_i is the number of the proximate ancestral vertex ($a_i < i$) and g_i is the age of vertex i, i.e. number of generations from this vertex up to the contemporary generation ($g_i \geq 0$ and $g_i \leq g_j$, if $j = a_i$ and $i > 1$). For the root vertex $a_1 = 0$, for any final vertex (leaf) i, g_i=0. We shall designate the sets of numbers a_i and g_i as {a} and {g}.

Let define μ(i/j) as the probability of a mutation from a nucleotide j to a nucleotide i per generation (Majewski, Ott, 2003). If μ(i/j) are given, a sequence of the root vertex uniquely determines the probability of sequences of all remaining vertices. The random mutation process causes the genetic distance between the leaves that is not certainly proportional to their distance in generations.

Because of that, a reconstruction of the origin tree based on similarity and difference between the leaves is a NP complete problem.

The full ETR model includes the number of ancestor vertices and the number of mutation events on the origin tree. The number of trees T_N rapidly increases with the number of leaves, *N*. Namely:

$T_{N=}[2(N\text{-}1)]!/[2^N(N\text{-}1)!]$.

For $N = 50$ this number approaches $3*10^{76}$.

Method, algorithm, implementation, and scenarios

We present a new ETR approach based both on the reduction the ETR problem to the known Traveling Salesman Problem (TSP) and on the minimization of the ETR model. The TSP solution is achieved by rearranging the leaf order to minimize the cycle length over the leaves (Korostensky, Gonnet, 2000). A powerful algorithm of TSP solution, utilizing the strength of the Guided Evolution Strategies (GES) method (Mester, Bräysy, 2004), was developed and adapted to the ETR problem. The TSP solution reduces the set of acceptable binary trees to at least *N*!/2. For the above example with $N = 50$, the number of trees decreases more than 10^{12} times. A single tree of the reduced set is selected by using the Average Linkage Clustering (ALC) method. The final tree, hereafter referred to as MBK tree, is achieved by combining neighboring vertices of the binary tree to minimize the number of parameters defined the ETR model (Korostishevsky *et al.*, 2001). The robustness and effectiveness of the model was established by simulation experiments. The resulted trees were compared with the trees obtained by the UPGMA algorithm (PHYLIP, 1995).

- The proposed algorithm includes the following functions:
- Simulating a tree of the given complexity.
- Simulating leaves of the given tree and the given root sequence.
- Evaluating ETR using the pairwise distance matrix between leaves.
- Estimating the quality and robustness of the ETR solution.

The software was built using C# language and Graphic User Interface (GUI) of Microsoft® Visual Studio.NET (http://msdn.microsoft.com/netframework/technologyinfo/howtoget/). The duration time on an ordinary computer, Pentium-4, is a few seconds for some hundreds of leaves. The simulation results on shaggy trees, $N > 100$, illustrate the effectiveness of the method.

Figure 1 is the program’s screenshot, followed by the description of the input stages and the output forms. The user interface was designed as a one-window form application with the control buttons on the left and the resulted trees on the right. Most frequent ETR scenarios include:

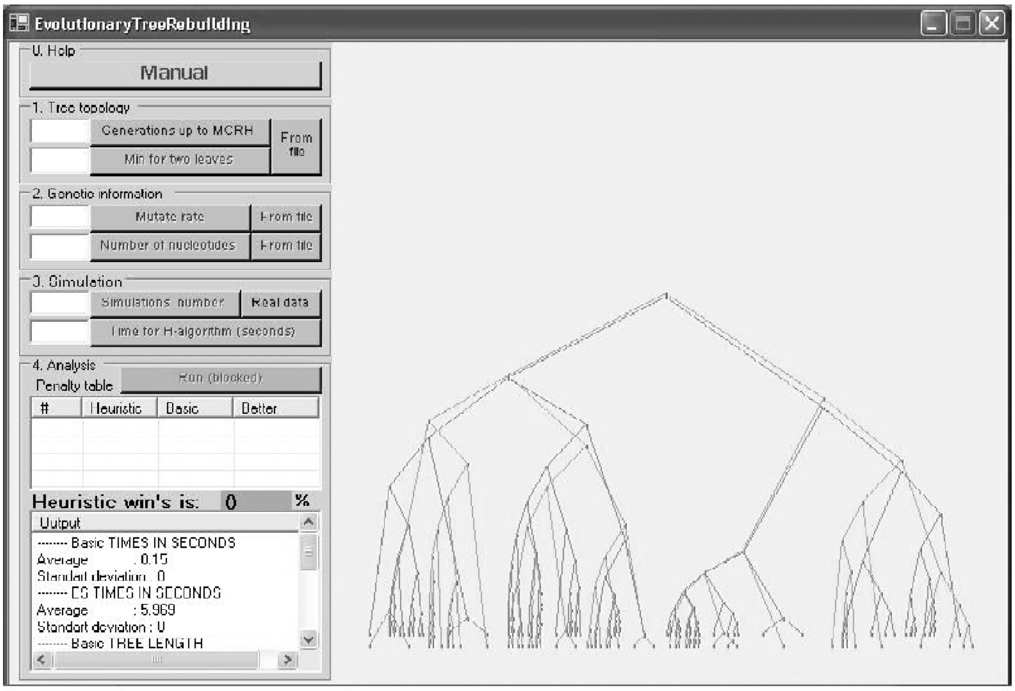
- The tree for haplotype simulations defined by user.
- The haplotypes loaded from user’s file.
- The haplotypes simulated on a random tree.

Simulation results and Discussion

Simulations were done on three different trees (with 19, 54 and 110 leaves).
Evaluation parameters: TH-tree height (1000, 10000), HL-haplotype length (50,100), MR-mutation rates (0.001, 0.0005). 1000 simulations were done for each of the 3*2*2*2=24 parameter combinations. The results are presented in Table and Figure 2.

Quality of the resulted trees was estimated by relative tree length, as a ratio reconstructed tree length / original tree length. The tree length is defined as one half of the minimal cyclic way through the leaves. This minimum way is achieved on original tree if the distances between leaves are proportional to age of MRCA. In our case, the distances between leaves are randomly deviating among 1000 simulations done for a given tree with fixed parameter values.

It can be easily seen that with the growing complexity of original tree our algorithm displays a growing advantage over the standard UPGMA algorithm.



#	Heuristic	Basic	Better

Heuristic win's is: 0 %

```
Output
-------- Basic TIMES IN SECONDS
Average          : 0.15
Standart deviation : 0
-------- ES TIMES IN SECONDS
Average          : 5.969
Standart deviation : 0
-------- Basic TREE LENGTH
```

Fig. 1. User interface.

Table. Simulation results for a tree with 110 leaves

Simulation parameters				Simulated tree		UPGMA		MBK tree	
Leaf number	Tree height	Mutation rate	Haplotype length	Tree length	STDEV	Tree length	STDEV	Tree length	STDEV
110	10000	0.0005	100	5573	1.9229	5524	1.8243	5259	1.9189
110	10000	0.0005	50.0	2786	1.2790	2730	1.2265	2538	1.2802
110	10000	0.0010	100	6947	1.7577	6802	1.5834	6503	1.6927
110	10000	0.0010	50.0	3472	1.3079	3313	1.0249	3104	2.6594
110	1000	0.0005	100	1326	1.5285	1336	1.5410	1233	1.5369
110	1000	0.0005	50.0	689.2	0.9809	691.2	0.9954	616.7	0.9848
110	1000	0.0010	100	2244	1.7744	2247	1.7687	2100	1.8093
110	1000	0.0010	50.0	1125	1.2490	1123	1.2653	1015	1.2421

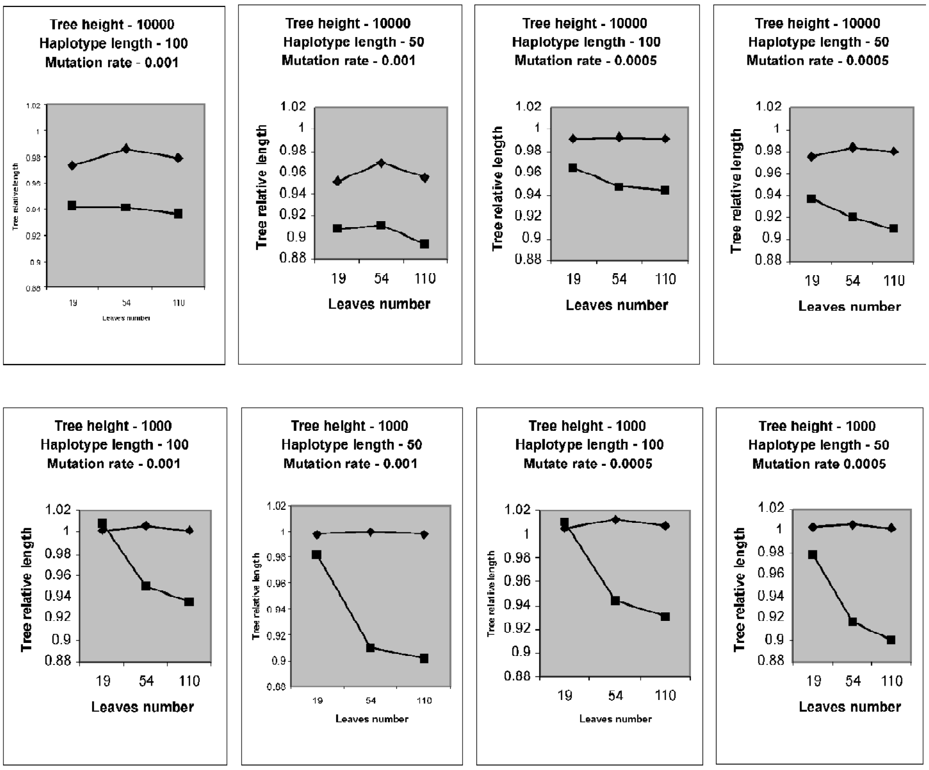


Fig. 2. Relative length of two reconstructed trees as a function of leaf number (squares for MBK and diamonds for UPGMA).

References

Bonné-Tamir B., Korostishevsky M., Redd A.J., Pel-Or Y., Kaplan M.E., Hammer M.F. Maternal and paternal lineages of the Samaritan isolate: mutation rates and time to most recent common male ancestor // Annals of Human Genetics. 2003. V. 67. P. 153–164.

Korostensky C., Gonnet G.H. Using traveling salesman problem algorithms for evolutionary tree construction // Bioinformatics. 2000. V. 16. P. 619–627.

Korostishevsky M., Ginzburg E., Bonné-Tamir B. Mutation origin reconstruction based on adjacent haplotypes // The First Workshop on Information Technologies Application to Problem of Biodiversity and Dynamics of Ecosystem in North Eurasia. Novosibirsk. Russia. 2001. Abstract P219.

Majewski J., Ott J. Amino acid substitutions in the human genome: evolutionary implications of single nucleotide polymorphisms // Gene. 2003. V. 305. P. 167–173.

Mester D., Bräysy O. Active guided evolution strategies for large-scale vehicle routing problems with time windows // Computers and Operations Res. 2004. (on line), 1–22.

PHYLIP. Phylogeny Inference Package, release 3.57c. University Washington, USA 1995. (http://evolution.genetics.washington.edu/phylip.html).

THE CONSERVATION OF TRANSCRIPTION FACTOR-BINDINS SITES IN *SACCHAROMYCES* GENOMES

Kovaleva G.Y.[*1,2], Mironov A.A.[1], Gelfand M.S.[2]*

[1] Moscow State University, Department of Bioengeneering and Bioinformatics, Moscow, Russia;
[2] Institute for Information Transmission Problems, RAS, Moscow, Russia
* Corresponding author: e-mail: kovaleva@iitp.ru

Keywords: *comparative genomics, yeast, Saccharomyces, transcription, regulatory site, Gcn4p, Met31/Met32, Cbf1/Met4/Met28, Leu3*

Resume

Motivation: two independent studies on prediction of regulatory sites in genomes of the genus *Saccharomyces* by comparative analysis produced contradictory results. In this study we studied the conservation rate of known and predicted binding sites for regulatory proteins of several metabolic pathways in *Saccharomyces* genomes.

Results: comparative analysis of seven *Saccharomyces* genomes shows that in most cases the binding sites are not perfectly conserved, or conserved but in a different position. This observation contradicts previously postulated statements that functionally significant regions are absolutely conserved and occupy the same in related genomes.

Introduction

Extracting the complete functional information encoded in a genome — including genic, regulatory and structural elements — is a central challenge in biological research. Prediction of non-protein-coding functional regions, such as regulatory elements, is especially difficult because of are usually short (6–15 bp for *S.cerevisiae* and many other eukaryotic genomes), often degenerate, and can reside on either strand of DNA at variable distances from the genes they control. Since functional sequences tend to be conserved through evolution, they can appear as 'phylogenetic footprints' in alignments of genome sequences of different species (Hardison *et al.*, 1997). Recently, two groups sequenced several *Saccharomyces* genomes. The main goal of these studies was to predict the regulatory sites in *Saccharomyces* spp. using multiple whole-genome alignment in one case (Kellis *et al.*, 2003) and multiple alignments of the gene upstream regions in another (Cliften *et al.*, 2003). Results were represented as two lists of predicted binding motifs. Our comparison of these lists shows a very moderate intersection even accounting to fact that in study by Kellis *et al.* the predicted motifs were constructed using IUB codes. This prompted us to analyze the conservation rate for known and predicted binding sites in *Saccharomyces* genomes in more detail.

Methods

Complete genome sequences of *Saccharomyces cerevisiae* and *Candida albicans* were extracted from GeneBank. The fragments covering 750 nucleotides of upstream regions and 150 nucleotides of protein-coding regions were considered. Search for orthologs was done using fungiBLAST (http://www.ncbi.nlm.nih.gov/BLAST/Genome/FungiBlast.html). In some cases orthologous regions for the protein-coding part of a used fragment were not found; and such fragments were ignored. On the contrary, regions upstream of orthologous genes were used even if they did not produce a strong alignment.

For identification of orthologous genes and site patterns, Genome Explorer program (Mironov *et al.*, 1999) was used. SignalX (Mironov *et al.*, 2000) was used to construct nucleotides weight matrices. Multiple sequence alignments were done using ClustalX (Thompson *et al.*, 1997).

Results and Discussion

Site conservation. Unlike many eukaryotic organisms yeasts are able to synthesize amino acids. Therefore, they have corresponding metabolic pathways and regulators. The global regulator of amino acid biosynthesis is Gcn4p. It translation is activated condition of starvation for some amino acids. The regulator binds to Gcn4p-responsive element (GRE), with the consensus TGACTC in the upstream regions of regulated genes (Natarajan *et al.*, 2001; Hinnebusch, Natarajan, 2002). It is also known that in upstream regions of genes regulated by Gcn4p usually contain at least two GRE. Based on published experimental data, we seleted nine genes that are certainly regulated by Gcn4p: *HIS3*, *ARG8*, *ARG1*, *ADE4*, *ILV1*, *TRP4*, *HIS4*, *HIS7*, and *ILV2* (Natarajan *et al.*, 2001). Using the TRANSFAC database we constucted a position weight matrice to identify significant binding sites in the upstream regions of these genes. The conservation rate for all these sites are shown in Table 1.

Table 1. Conservation rates of known, strong and weak predicted binding sites of Gcn4p in 5 of 7 analyzed *Saccharomyces* genomes. The number of conserved sites in each genome is set in bold. The numbers in parentheses are the number of sites that could not be conserved since orthologous region is absent or not sequenced in corresponding genome

Genome/conserved sites	Known	Predicted	Weak
S. cerevisiae	**11**	**6**	**32**
S. paradoxus	**10**	**4**	**13**
Non-conserved sites	1 (0)	2 (0)	19 (0)
S. mikatae	**6**	**4**	**10**
Non-conserved sites	4 (1)	2 (0)	19 (3)
S. kudriavzevii	**3**	**3**	**5**
Non-conserved sites	1 (7)	0 (3)	7 (20)
S. bayanus	**0**	**2**	**2**
Non-conserved sites	0 (11)	0 (4)	7 (23)

From these data it follows that the conservation rate of known and strong predicted binding sites is similar. Still, even strongest sites are not necessary conserved in all examined genomes, as it has been thought previously.

The biosynthesis of methionine is regulated by three more or less independent regulators/regulatory complexes: Gcn4p, Met31/Met32, and Cbfl/Met4/Met28. Prediction of binding sites for Gcn4p was done using the same matrice as above. For the other two regulatory complexes, only the binding site consensus is known: TCACGTG for the Met31/Met32 complex and AAACTGTGG for the Cbfl/Met4/Met28 complex (Thomas, Surdin-Kerjan, 1997). Nevertheless, we constructed the corresponding matrices and applied them to known regulatory genes (Thomas, Surdin-Kerjan, 1997). Thus subsets of known and predicted binding sites were identified. As it can be seen in Table 2, the conservation rates of these groups of sites are similar to those for Gcn4p (see above).

The leucine biosynthesis pathway is regulated by two regulators Gcn4p, a master regulator of amino acid biosynthesis, and specific regulator Leu3. Leu3 binding sites are not given in TRANFAC database, and we constructed a position weight matrice on the basis of the consensus (Kohlhaw, 2003) and few experimentally known binding sites (most genes of this regulon are known to be regulated from expression array experiments).

The conservation rates for these regulators are given in Table 3 and are more or less similar to the rates for other studied regulators. However, the binding sites for Leu3, both known and predicted, seem to be more conserved compared to the first two cases. The difference can be due to the lengh of the signal: for Leu3 it seems to be about 10 nucleotides, as internal positions also contribute to the strength of interaction.

Thus, the standard technique of comparative genomics can not be used to yeast genomes without corrections. There is no universal conservation of binding sites that is often observed in prokatyotic organisms.

Table 2. Conservation rates of known, strong and weak predicted binding sites of Met31/Met32 and Cbf1/Met4/Met28 complexes in five of seven analyzed *Saccharomyces* genomes. Notations as in Table 1

Genome/conserved sites	Met31+ Met32	Known	Predicted	Weak	Cbf1-complex	Known	Predicted	Weak
S. cerevisiae		**10**	**11**	**80**		**8**	**15**	**51**
S. paradoxus		**8**	**9**	**22**		**2**	**6**	**16**
Non-conserved sites		0 (2)	1 (1)	43 (5)		5 (1)	6 (3)	35 (0)
S. mikatae		**4**	**3**	**7**		**2**	**2**	**8**
Non-conserved sites		2 (4)	2 (6)	22 (51)		2 (4)	2 (11)	20 (23)
S. kudriavzevii		**2**	**1**	**2**		**2**	**0**	**4**
Non-conserved sites		1 (7)	2 (8)	10 (68)		2 (4)	2 (13)	11 (36)
S. bayanus		**2**	**2**	**5**		**1**	**1**	**0**
Non-conserved sites		0 (8)	2 (7)	22 (53)		3 (4)	1 (13)	18 (33)

Table 3. Conservation rates of known, strong and weak predicted binding sites of Leu3 in five of seven analyzed *Saccharomyces* genomes. Notations as Tables 1 and 2. Fractional numbers reflect that strong (known or predicted) sites may become weaker, but still above the threshold, due to partially inactivating mutations

Genomes	Known	Predicted	Weak
S. cerevisiae	5	6	12
S. paradoxus	**4**	**5.5**	**5**
Non-conserved sites	0 (1)	0.5 (0)	7 (0)
S. mikatae	**4**	**4**	**1**
Non-conserved sites	0 (1)	1 (1)	6 (5)
S. kudriavzevii	**1.5**	**3**	**1**
Non-conserved sites	0.5 (3)	0 (3)	5 (7)
S. bayanus	**2.5**	**1**	**1**
Non-conserved sites	1.5 (1)	0 (5)	5 (7)

Site clusterization. Its is known that most eukatyotic regulatory signals are short and the selectivity of regulator binding is obtained by clustering of sites them in the upstream regulatory regions. We considered clusterization of binding sites of Gcn4p in the upstream regions of nine genes listed above of *Saccharomyces cerevisiae* and *Candida albicans* genomes. Clusterization of binding sites is not as strict criterion as absolute conservation of the exact motif at a certain position. However, our data shows that clusterization of binding sites is evolutionary significant because of it is observed in so divergent genomes. Still, the statistical parameters are unique for every cluster in each genome. We observed that genes known to be regulated by Gcn4p in most cases have clusters of candidate Gcn4p binding sites in the upstream regions genes from *S.cerevisiae* and from *C. albicans*.

Thus, experimentally known and strong predicted binding sites for transcriptional regulators of several yeast metabolic pathways have a similar conservation rate in related genomes, whereas weak predicted sites are less conserved. However, even in closely related genomes conservation of experimentally defined sites is not guaranteed. A weaker form of comparative analysis, requiring conservation of site clusters irrespective of their parameters, can be used at larger evolutionary distances, but the recognition rules turn out to be not very specific.

Acknowlegements

We are grateful to Lena Stavrovskaya for kindly provided software. This study was supported by grants from HHMI, LICR, RFBR, the Fund for Support of Russian Sciences, and the Program in Molecular and Cellular Biologyof RAS.

References

Cliften P., Sudarsanam P., Desikan A., Fulton L., Fulton B., Majors J., Waterston R., Cohen B.A., Johnston M. Finding functional features in Saccharomyces genomes by phylogenetic footprinting // Science. 2003. V. 301. P. 71–76.

Hardison R.C., Oeltjen J., Miller W. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome // Genome Res. 1997. V. 7. P. 959–966.

Hinnebusch A.G., Natarajan K. Gcn4p, a master regulator of gene expression, is controlled at multiple levels by diverse signals of starvation and stress // Eukaryotic Cell. 2002. V. 1. P. 22–32.

Kellis M., Patterson N., Endrizzi M., Birren B., Lander E.S. Sequencing and comparison of yeast species to identify genes and regulatory elements // Nature. 2003. V. 423. P. 241–254.

Kohlhaw G.B. Leucine biosynthesis in fungi: entering metabolism through the back door // Microbiol. Mol. Biol. Rev. 2003. V. 67. P. 1–15.

Mironov A.A., Koonin E.V., Roytberg M.A., Gelfand M.S. Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes // Nucleic Acids Res. 1999. V. 27. P. 2981–9.

Mironov A.A., Vinokurova N.P., Gelfand M.S. GenomeExplorer: software for analysis of complete bacterial genomes // Mol. Biol. 2000. V. 34. P. 222–231.

Natarajan K., Meyer M.R., Jackson B.M., Slade D., Roberts C., Hinnebusch A.G., Marton M.J. Transcriptional profiling shows that Gcn4p is a master regulator of gene expression during amino acid starvation in yeast // Mol. Cell Biol. 2001. V. 21. P. 4347–4368.

Thomas D., Surdin-Kerjan Y. Metabolism of sulfur amino acids in Saccharomyces cerevisiae // Microbiol. Mol. Biol. Rev. 1997. V. 61. P. 503–32.

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

THE DISTANCE FUNCTION FOR COMPUTING THE CONTINUOUS DISTANCE OF BIOPOLYMER SEQUENCES

Margaryan T.V., Hakobyan G.O.*

Dept. of Physics, Yerevan State Univ., Yerevan 375025, Armenia
* Corresponding author: e-mail: gaghakob@ysu.am

Keywords: *amino acid sequence, distance matrix, distance function*

Summary

Motivation: Character strings representing the monomer succession in linear polymers are being scrutinized from many different semantic aspects. However, as any abstraction, sequence perception preferentially by means of discrete residue algorithms necessarily remains one-sided. It doesn't facilitate the elucidation of those biologically meaningful molecular traits, which depend, generally speaking, on a biopolymer's chain continuity and integrity. To widen the scope of bioinformatics towards a desirably more holistic compilation of both the intrinsic molecular properties and some superordinate correlations in the living world, it is legitimate to ask in how far the use of *functions in the mathematical sense of the word* may assist researchers with addressing sequence-depending problems better than discrete maths does to date. Naturally, from this method of comparison the loss of information is excepted. To the contrary received more fine results (the practical results obtained by this approach for groups of 99 proteins from SwissProt –web location srs6.ebi.ac.uk (Heymann *et al.*, 2002).

Nevertheless, none of the global, local or multi-local sequence comparison/alignment algorithms yields a similarity measure (score) that allowed a researcher to interrelate results from run to run. Indeed, whatever common method a researcher applies to compare sequence A to B, A to C – the according scores do not allow him to judge how similar or unalike B will be with respect to C. The simple explanation of this unsatisfying situation is the lack of *metrical* distances, esp. the non-compliance of the triangle inequation for the usual scores attributed to any three sequences pairs.

Results: We construct a distance function with the help of the given distance matrix. Being itself as a continuous function it keeps all the properties which has the given distance matrix. In particular, we proved that if the given matrix satisfies the condition of triangle inequality then the constructed distance function must also satisfy this condition.

Introduction

In S.M. Ulam's monography (Ulam, 1972) the evolutionary distance of any two words was defined as the minimal sum of operation costs (for insertion, deletion, mutation, resp.) necessary for transforming word *a* stepwise into word *b* (commonly known as *minimum edit distance*).

In 1970, S.B. Needleman, C.D. Wunsch (Needleman, Wunsch, 1970) introduced a dynamic programming algorithm for the global similarity assessment/alignment of two words, with a constant gap penalty $g(k) = \alpha$ for all positions *k* opposite to an $\varnothing$. Importantly, insertions and deletions are not allowed to be adjacent. Dynamic programming was further developed by D. Sankoff (see Sankoff, 1972) and independently applied by P.H. Sellers (see Sellers, 1974) and R.A. Wagner, M.J. Fisher (Wagner, Fischer, 1974), to determine the distance of real amino acid successions.

The history of functional ("continuous") distances of two numerical successions, the elements of which are "letters" in a numerical alphabet *V* with a given distance matrix *D*, started evidently from papers by L.V. Kantorovich, G.S. Rubinstein (Kantorovich, Rubinstein, 1957) and by L.N. Wasserstein (Wasserstein, 1969). In the book D. Sankoff, J.B. Kruskal (1999) papers dedicated to the method of *Time Warping* are summarized.

The "Metric" method is an approach of transition the discrete sequences to "continuous" functions and definition of distance of two strings by this functions (Heymann *et al.,* 2002; Heymann *et al.*, 2004). The distance between two sequences $a = a_1a_2...a_n$ and $b = b_1b_2...b_m$ is classically defined by the minimum sum of distances of the characters $d(a_i,b_j)$, $d(a_i,\varnothing)$, $d(\varnothing,b_j)$ where a_i, b_j are some characters, and $\varnothing$ is the "empty" character $(1 \le i \le n, 1 \le j \le m)$, where the minimum is extended over all alignments of *a* with *b*. In this definition, characters are taken as isolated elements in a sequences and no neighbourhood influence on a_i, b_j or $\varnothing$ is being considered. In contrast, if a and b were brought in accordance to functions $a(t)$ and $b(t)$ of a continuous argument *t*, one observes for fixed argument values t_0 a "distance" between $a(t_0)$ and $b(t_0)$ that is influenced by function values around t_0.

The transition into the metrical space and the deduction mode of a "continuous" distance between two amino acid sequences is a process embracing several steps: the transition from a character alphabet to a numerical, the dediscretization of sequences, the dediscretization of distance matrix, defines the distance functional $d(a,b) = c(m,n) \cdot \inf_{\varphi,\psi} \int_0^T D[a(\varphi(t)), b(\psi(t))]\, \lambda_{\varphi\psi}(t)dt$, where the lower boundary is chosen from all allowed φ, ψ, $\lambda_{\varphi\psi}$; $c(m,n) = \frac{m+n-2s}{m+n}$, s is the length of the longest coinciding segment in *a* and *b*. The inner functional terms $\{\varphi(t),\ \psi(t),\ \lambda_{\varphi\psi}(t)\}$ is constructed acc. to an algorithm, which ensures a value of the integral in the right-hand part of the above relation "near" to the minimum value (Heymann *et al.*, 2002; Heymann *et al.*, 2004).

The central role in this approach plays the distance function of two independent variables independent of the way of introducing the metrics in the space of continuous function. As none of the "distance" matrices of amino acids known to us fulfills one of the axioms of metricity, namely the triangle inequation, we use the terms "metrics" and "distance" in quotation marks. Anyhow, in the result of a minor correction any of the "distance" matrices in use will fulfill the triangle inequation. As we do not know the biological consequences of such corrections we will deal with uncorrected matrices, although lateron we will give examples for computed distances of amino acid sequences on the basis of corrected matrices (we use the distance matrix MIATA).

In Heymann *et al.*, 2002; Heymann *et al.*, 2004 the constructed distance function does not satisfy inequalities of a triangle, even then when the distance matrix through which is constructed it function, will satisfy inequalities of a triangle.

Methods

In this paper we have used the methods of mathematical analysis.

Results and Discussion

We construct a continue distance function of two independent variables, with the help of the given distance matrix. We proved that if the given matrix satisfies the condition of triangle inequality then the constructed distance function must also satisfy this condition.

It is interesting to note that in all known to us matrixes those amino acids don't satisfy the triangle inequality, to which different triplets correspond. We don't know how these matrixes have been construct, but we think that they can be charged a little such that their elements satisfy the triangle inequality. It is possible to reach, in our opinion, if initially the distances between amino acids will be defined such that those satisfy the triangle inequality.

Acknowledgements

The authors are thankful to referee's for useful comments and criticism.

References

Heymann S., Gabrielyan O., Ghazaryan H.G., Danielyan E., Hakobyan G.G., Hakobyan G.H. A metrical space of biological sequences // Proc. of the ISAAC Conference on Analysis, Yerevan, Armenia, Sept. 17-21, 2002. P. 1–18.

Heymann S., Gabrielyan O., Ghazaryan G., Danielyan E., Hakobyan G.G., Hakobyan G.H. Towards a metrical space of biological sequences // 3-th International Conference BGRS'2002. Novosibirsk, 2002. Russian.

Kantorovich L.V., Rubinstein G.S. On a function space and certain extremum probleme // Dokl. Akad. Nauk SSSR. 1957. V. 115, N 5. P. 1058–1061.

Needleman S.B., Wunsch C.D. A general method applicable to the search for similarities in the amino acid sequences of two proteins // J. Mol. Biol. 1970. V. 48. P. 443–453.

Sankoff D. Matching sequences under deletion – insertion constraints // Proc. Natl Acad. Sci. USA. 1972. V. 68. P. 4–6.

Sankoff D., Kruskal J.B. Time Warps, String Edits and Macromolecules. CSLI Publications, ISBN 1-57586-217-4 (originally published 1983 by Addison-Wisley, Reading, MAS) 1999.

Sellers P.H. An algorithm for the distance between two finite sequences // J. Combinator Theor. A 1974. V. 16. P. 253–258.

Ulam S.M. Some combinatorial problems studied experimentally on computing machines. S.K. Zaremba, New-York: Academic Press. 1972.

Wagner R.A., Fischer M.J. The string-to-string correction problem // J. Assoc. Comput. Mach. 1974. V. 21. P. 168–173.

Wasserstein L.N. Markov processes over denumerable products of spaces describing Large Systems of automata // Problems of Information Transission 1969. 5, 47-52.

COMPARATIVE COMPLETE GENOME SEQUENCE ANALYSIS OF *CORYNEBACTERIA*

***Nishio Y.*[1]*, *Nakamura Y.*[2], *Usuda Y.*[1], *Kawarabayasi Y.*[3], *Yamagishi A.*[4], *Kimura E.*[1], *Matsui K.*[1], *Sugimoto S.*[5], *Kikuchi H.*[6], *Ikeo K.*[2], *Gojobori T.*[2]**

[1] Institute of Life Sciences, Ajinomoto Co., Inc., Kawasaki, Japan; [2] Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Mishima, Japan; [3] Research Center for Glycoscience, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan; [4] Department of Molecular Biology, Tokyo University of Pharmacy and Life Science, Hachioji, Japan; [5] Fermentation & Biotechnology Laboratories, Ajinomoto Co., Inc., Kawasaki, Japan; [6] National Institute of Technology and Evaluation, Shibuya , Japan

* Corresponding author: e-mail: yousuke_nishio@ajinomoto.com

Keywords: *evolution, protein thermostability, Corynebacterium, comparative genomics*

Summary

Motivation: The three corynebacterial complete genomes have been sequenced. Comparative genome sequence analysis will highlight the evolutional history responsible for their functional differentiation.

Results: Corynebacterium efficiens but not *Corynebacterium glutamicum* can grow above 40 °C. The genome GC contents of *C. efficiens* was 63.1 % and about 10 % higher than that of *C. glutamicum* or *Corynebacterium diphtheriae*. This difference was reflected in codon usage and nucleotide substitutions. By analyzing orthologous gene pairs with 60–95 % amino acid sequence identity between *C. efficiens* and *C. glutamicum*, we observed tremendous bias in amino acid substitutions. Especially, three asymmetrical amino acid substitutions: from lysine to arginine, serine to alanine and serine to threonine, were important for protein thermostability in *C. efficiens*. The accumulations of these amino acid substitutions were responsible for thermostability in *C. efficiens* and greater GC contents.

Introduction

Monosodium glutamate has been known to be the umami substance which improves the taste of foods. *C. glutamicum* is a well known for industrial glutamic acid production by fermentation. *C. efficiens* has been shown to be a near relative of *C. glutamicum*, and both of them have been recognized as glutamic acid producing species. The optimal temperature for glutamate production by *C. glutamicum* is around 30 °C, and this microorganism can neither grow nor produce glutamate at 40 °C or above. On the other hand, *C. efficiens* can grow and produce glutamate above 40 °C. The thermostability of *C. efficiens* is a useful trait from an industrial viewpoint as it reduces the considerable cost of cooling needed to dissipate the heat generated during glutamate fermentation.

The finding for the difference of growth temperature and GC contents between *C. efficiens* and *C. glutamicum* (Fudou *et al.*, 2002), and the more protein thermostability of *C. efficiens* compared to that of *C. glutamicum* (Kimura *et al.*) provides an attractive topic for the comparative genomics. Furthermore, because these two species are phylogenetically closely related species, we can compare more than 1,000 orthologous genes with 60–95 % amino acid sequence identity one by one. This is the greatest advantage of this study. In previous study for genome-wide comparisons between thermophilic archae and mesophilic bacteria, amino acid residues do not correspond one-to-one. We have tried here to elucidate the mechanism underlying the thermal stability of *C. efficiens* by a genome-wide comparison of amino acid substitutions, in the hope that such a comparison may indicate a general method for protein thermostabilization.

Method

The complete genome sequences of *C. glutamicum* (Ikeda, Nakagawa, 2003), *C. efficiens* (Nishio *et al*., 2003) and *C. diphtheriae* (Cerdeno-Tarraga *et al*., 2003) were obtained from DDBJ/EMBL/GenBank. Orthologous genes are defined as the best pair of homologues in comparisons between two organisms (Tatusov *et al*., 1997).

Results and Discussion

The features of three *Corynebacteria* were summarized on Table 1. To gain an overview of corynebacterial genome structure, we compared the GC contents profile. *C. glutamicum* had a GC content between 50 % and 60 % in most regions of the chromosome, and its average GC content was 53.8 %. On the other hand, the average GC content of *C. efficiens* was 63.4 %, higher than *C. glutamicum* over the entire chromosome. *C. diphtheriae* was used as an outgroup of *C. efficiens* and *C. glutamicum*. *C. diphtheriae* showed a window analysis profile of GC content more similar to *C. glutamicum* than to *C. efficiens*. This suggests that the ancestral genome structure of *Corynebacteria* may be closer to that of *C. glutamicum* than to that of *C. efficiens,* and the thermostability of *C. efficiens* may be acquired after divergence from the common ancestor of *C. glutamicum* and *C. efficiens*.

Table 1. Summary of characteristics of *Corynebacteria*

	C. efficiens	*C. glutamicum*	*C. diphtheriae*
Upper temperature limit for growth (℃)	45	40	-
Glutamate production at 32 ℃ (%)*	80	100	-
Glutamate production at 37 ℃ (%)	78	40	-
Genome size (bp)	3,147,090	3,309,401	2,488,635
GC content (%)	63.1	53.8	53.5
Number of predicted gene	2,942	3,099	2,320

* Glutamate production in typical experiments using the biotin limitation method as a percent of the production by *C. glutamicum* at 32 °C.

To estimate the mutation responsible for the protein thermostability in *C. efficiens*, we analyzed asymmetrical amino acid substitutions between *C. efficiens* and *C. glutamicum.* The orthologous ORFs with identity of amino acids sequence under 60 % were omitted from the analysis, because of the large calculated p-distance value of 0.4 and the need to take account of backward and parallel mutations (Nei, Sudhir 2000). 1,619 orthologous pairs of genes with identity from 60 % to 95 % (p-distance value 0.2) were used to examine position-specific mutations. The most frequently observed asymmetrical amino acid substitutions between *C. glutamicum* and *C. efficiens* was from Lys in *C. glutamicum* to Arg in *C. efficiens*. The following ones were Ser to Ala, Ser to Thr and Ile to Val. Some of the amino acid substitutions in this table have often been observed before in nature, with Leu, Ile, Val, and Met replacing each other. Because the fourth most frequent substitution, from Ile to Val, is commonly observed in situations unrelated to thermostabilization, the three most frequent substitutions (Lys to Arg, Ser to Thr, Ser to Ala) were assumed to be specific amino acid substitution pattern between *C. efficiens* and *C. glutamicum*, and the best candidates for stabilizing the proteins. Indeed many studies have suggested that the Lys to Arg substitution affects thermal stability (Vieille, Zeikus, 2001). If the evolutionary development of the thermal stability of proteins is responsible for the thermostability of *C. efficiens* itself, then the observed amino acid substitutions must be adaptive mutations leading to overall thermostability. In a separate study, the thermal stability of 13 pairs of enzymes on the glutamic acid and lysine biosynthetic pathways in the two species were compared on the basis of the enzymatic activities remaining after heat treatment of crude extracts (Kimura *et al.*). In Table 2 the numbers of the three kinds of amino acid substitutions within the amino acid sequence are assigned points depending on their directions, and we compare the number of calculated points with the experimental results of enzyme thermal stability. Nine out of 13 enzymes, the thermostabilities of which had been measured, agree with the calculated points, 3 can not be determined, and only one does not coincide (Table 2). These results suggest that there is a significant correlation between the three kinds of amino acid substitution and the thermal stability of proteins.

Table 2. Check of predictions against actual measurements

Entry	Enzyme	Thermostable species	Point	Result
1	2-Oxoglutarate dehydrogenase	*C. efficiens*	0	-
2	Glutamate dehydrogenase	*C. efficiens*	1	Yes
3	Isocitrate lyase	*C. efficiens*	2	Yes
4	Phosphofructokinase	*C. efficiens*	-3	No
5	Fructose-1-phosphate kinase	*C. efficiens*	5	Yes
6	Isocitrate dehydrogenase	*C. efficiens*	4	Yes
7	Aconitase	*C. efficiens*	0	-
8	Phosphoenolpyruvate carboxylase	*C. efficiens*	10	Yes
9	Citrate synthase	*C. efficiens*	2	Yes
10	Aspartate kinase	*C. glutamicum*	-1	Yes
11	Dihydrodipicolinate synthase	*C. efficiens*	0	-
12	Diaminopimelate dehydrogenase	*C. glutamicum*	-2	Yes
13	Diaminopimelate decarboxylase	*C. efficiens*	2	Yes

Point is defined as the difference between the sum of the three kinds of substitutions from *C. glutamicum* to *C. efficiens* (Lys to Arg, Ser to Ala and Ser to Thr) and the sum of the reverse substitutions (Point = $\{\text{number of } (\text{Lys}\rightarrow\text{Arg} + \text{Ser}\rightarrow\text{Ala} + \text{Ser}\rightarrow\text{Thr})\} - \{\text{number of } (\text{Arg}\rightarrow\text{Lys} + \text{Ala}\rightarrow\text{Ser} + \text{Thr}\rightarrow\text{Ser})\}$).

Results are indicated by 1) Yes: when the enzyme from *C. efficiens* was more thermostable and the point is positive, or when the enzyme from *C. glutamicum* was more thermostable and point is negative. 2) -: when the point was zero. 3) No: all other cases.

References

Cerdeno-Tarraga A.M. *et al.* The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129 // Nucleic Acids Res. 2003. V. 31. P. 6516–6523.

Fudou R. *et al. Corynebacterium efficiens* sp. nov., a glutamic-acid-producing species from soil and vegetables // Int. J. Syst. Evol. Microbiol. 2002. V. 52. P. 1127–1131.

Kimura E. *et al.* manuscript in preparation.

Ikeda M., Nakagawa S. The *Corynebacterium glutamicum* genome: features and impacts on biotechnological processes // Appl. Microbiol. Biotechnol. 2003. V. 62. P. 99–109.

Nishio Y. *et al.* Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens* // Genome Res. 2003. V. 13. P. 1572–1579.

Nei M., Sudhir K. Molecular Evolution and Phylogenetics. Oxford University Press, New York, 2000. P. 17–31.

Tatusov R. L. *et al.* A genomic perspective on protein families // Science. 1997. V. 278. P. 631–637.

Vieille C., Zeikus G.J. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability // Microbiol. Mol. Biol. Rev. 2001. V. 65. P. 1–43.

NEW LTR RETROTRANSPOSABLE ELEMENTS FROM THE EUKARYOTIC GENOMES

Novikova O.*[1], Fursov M.[2], Beresikov E.[3], Blinov A.[1]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Novosibirsk Center of Information Technologies "UniPro", Novosibirsk, Russia; [3] Hubrecht Laboratory, Netherlands Institute for Developmental Biology, Utrecht, The Netherlands
* Corresponding author: e-mail: novikova@bionet.nsc.ru

Keywords: *mobile elements, LTR retrotransposons, distribution, evolution, computer analysis*

Summary

Motivation: Retrotransposons are a common class found in the eukaryotic genomes. LTR retrotransposons are classified into two main families, Pseudoviridae and Metaviridae. However, the taxonomic classification of LTR retrotransposons is likely more complecated due to the recent identification of new subgroups, such as Bel, and DIRS. Use of data generated by different genome sequencing projects to identify new LTR retrotransposons can give insights into the evolution and classification of this group of mobile elements.

Results: The distribution of LTR retrotransposable elements was analyzed from 38 eukaryotic genomes. In all 27 new elements were revealed in the 23 genomes investigated. The new elements fall into four known groups of LTR retrotransposons: Ty1/copia, 3; Bel , 3; DIRS, 2 and Ty3/gypsy, 18 new elements.

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

Introduction

Retrotransposable elements were classified into two large groups, long-terminal repeat (LTR) and non-LTR retrotransposons, on the basis of their overall structures. Recent phylogenetic analyses based on the amino acid sequences of Pol proteins demonstrated that each of these two groups is composed of several distinct clades, members of which are thought to be closely related in evolutionary terms (Malik *et al*., 1999; Malik, Eickbush, 1999). LTR retrotransposons were divided into two main families based on their structure, Pseudoviridae, which includes the Ty1/copia group and the caulimoviruses, and Metaviridae, which includes the Ty3/gypsy group and the retroviruses. However, recent studies revealed three new families: Bel, Xena and DIRS. Each family has a distinctive structural organization of open reading frames (ORF) and enzymatic modules.

A phylogenetic analysis of retrotransposons was initially based on reverse transcriptase (RT) sequences and it was revealed that LTR retrotransposons are younger than non-LTR retrotransposons and their distribution confirms this (Xiong, Eickbush, 1990). Subsequent reports have indicated that the LTR retrotransposon RT domains are the most divergent of all the elements and use of only this domain for tracing LTR retrotransposons phylogeny is inadequate sufficient. Therefore, integrase (INT) and ribonuclease H (RNH) were used later in phylogenetic analysis of LTR retrotransposons (Malik, Eickbush, 1999; Malik, Eickbuch, 2001).

In the present investigation, we analyzed 38 different eukaryotic genomes from sequencing projects by time-consuming HMM software.

Methods

Sequences of genomes were obtained from the appropriate database. A six-frame translation of all the sequences was produced. The alignments obtained by the CLUSTAL W software (Thompson *et al.*, 1994) were used as subsets in construction of profiles for further analyses. A total number of four alignments were obtained: 1) RT domains from described elements which are related to both Ty1/copia and Ty3/gypsy groups; 2) RT, RNH and INT domains from known Ty3/gypsy-like elements and retroviruses; 3) RT, Peptidase and INT domains from elements, which belong to the Bel group of elements; 4) RT domain from Ty3/gypsy-like and DIRS-like elements.

The HMMER 2.1.1 (http://hmmer.wustl.edu/) software was used to identify all the sequences in these subsets matching the model of the appropriate domains. Only matches with scores above zero were considered in the further analyses.

The results of HMMER searches were analyzed using the specially designed scripts that grouped into families and classified on the basis of their similarities to known retrotransposons, filtered based on the presence of appropriate domains and number of stop-codons.

Phylogenetic trees were generated by the neighbor-Joining method using the MEGA2 software package (Kumar *et al.*, 2001).

Results and Discussion

The list of observed genomes and the results of our searches are summarized in Table.

Elements from the Ty1/copia group can be divided into two phylogenetic clades: copia and Ty1. The elements of this group are spread throughout the genomes of fungi, plants and insects. We revealed three new elements from this clade: whiterot75 and magrapor298 from fungi *Phanerochaete chrysosporium* and *Magnaporthe grisea,* respectively, and aegypti1369 from mosquito *Aedes aegypti*.

In addition to known Bel-like elements from the genomes of invertebrates and vertebrates, such as *Caenorhabditis elegans* (Cer11 and Cer13), *Ascaris lumbricoides* (TAS), *Drosophila melano gaster* (MAX, Batumi, Bel and Roo), *Anopheles gambiae* (moose), *Danio rerio* (Catch1,2,3) and *Takifugu rubripes* (Catch), we obtained three new elements: brugia1 from *Brugia malayi*, cbrigg154 from *Caenorhabditis briggsae,* and ciona83 from the genome of *Ciona intestinalis*. According to the phylogenetic tree, this group of elements can be subdivided into two clades: Cer11-like elements and moose-like elements.

Only two new elements of the DIRS family were found in the genomes of *C. briggsae* (cbrigg98) and *Leishmania major* (leish1). This group of elements is the smallest of all the known LTR retrotransposons groups.

Eight clades were selected within the Ty3/gypsy group (Malik, Eickbush, 1999). The elements of this group were identified in the genomes of many organisms investigated. In our analyses, 19 new elements were revealed from 10 genomes: (i) fungi *Aspergillus fumigatus* (asperfum1), *Aspergillus nidulans* (anidulans1, 20), *P. chrysosoporium* (whiterot1, 37), *Neurospora crassa* (neuro1), *Candida albicans* (candida3); (ii) insects *A. aegypti* (aegypti76, 92, 93), *A. gambiae* (anoph130, 502, 961, 2539); (iii) nematoda *C. briggsae* (cbrigg46, 105); (iv) chordata *C. intestinalis* (ciona3, 16) and *D. rerio* (danio15) (Fig.).

After the addition of new elements, we can distinguish three new clades: Boudicca, Ylt1 and woot, but no Athila clade that has been selected previously (Malik, Eickbush, 1999). Most elements in new clades Ylt1 and woot were described in this investigation. It is clear that further analyses will allow to give more grounded conclusions about the existence of these clades. No LTR retrotransposons have been revealed in 15 of the 38 genomes investigated. These organisms belong to the most ancient eukaryotic taxa that confirm the more recent origin of LTR retrotransposons in comparison with non-LTR retrotransposons.

Table. Distribution of four LTR retrotransposon groups among 38 investigated genomes

Phylum	Species	Ty1/copia	Bel	DIRS	Ty3/gypsy	New elements
DIPLOMONADIDA	*Giardia lamblia*	.	.	.	.	
APICOMPLEXA	*Eimeria tenelli*	.	.	.	.	
	Plasmodium berghei	.	.	.	.	
	Plasmodium chabaudi	.	.	.	.	
	Plasmodium falciparum	.	.	.	.	
	Plasmodium knowlesi	.	.	.	.	
	Plasmodium vivax	.	.	.	.	
	Plasmodium yoelii	.	.	.	.	
	Theileria annulata	.	.	.	.	
EUGLENOZOA						
Kinetoplastida	*Trypanosoma brucei*	.	.	.	.	
	Leishmania major	.	.	+	.	+
CRYPTOPHYTA	*Guillardia theta*	.	.	.	.	
VIRIDIPLANTAE	*Triticum aestivum*	+	.	.	+	
	Arabidopsis thaliana	+	.	.	+	
	Oryza sativa	+	.	.	+	
ENTAMOEBIDAE	*Entamoeba histolytica*	.	.	.	.	
	Entamoeba invadens	.	.	.	.	
MYCETOZOA	*Dictyostelium discoideum*	.	.	+	+	
FUNGI						
Basidiomycota	*Phanerochaete chrysosporium*	+	.	.	+	+
	Cryptococcus neoformans	.	.	.	.	
Ascomycota	*Aspergillus fumigatus*	.	.	.	+	+
	Aspergillus nidulans	.	.	.	+	+
	Candida albicans	+	.	.	+	+
	Magnaporthe grisea	+	.	.	+	
	Neurospora crassa	.	.	.	+	+
	Saccharomyces cerevisiae	+	.	.	+	
	Schizosaccharomyces pombe	+	.	.	+	
Microsporidia	*Encephalitozoon cuniculi*	.	.	.	.	
METAZOA						
Nematoda	*Caenorhabditis elegans*	.	+	.	+	
	Caenorhabditis briggsae	.	+	+	+	+
	Brugia malayi	.	+	.	+	+
Platyhelminthes	*Schistosoma mansoni*	.	.	.	+	
Insecta	*Aedes aegypti*	+	+	.	+	+
	Anopheles gambiae	+	+	.	+	+
Chordata						
Urochordata	*Ciona intestinalis*	.	+	.	+	+
Teleostei	*Danio rerio*	+	+	+	+	+
	Tetraodon nigroviridis	.	+	+	+	
	Takifugu rubripes	+	+	.	+	

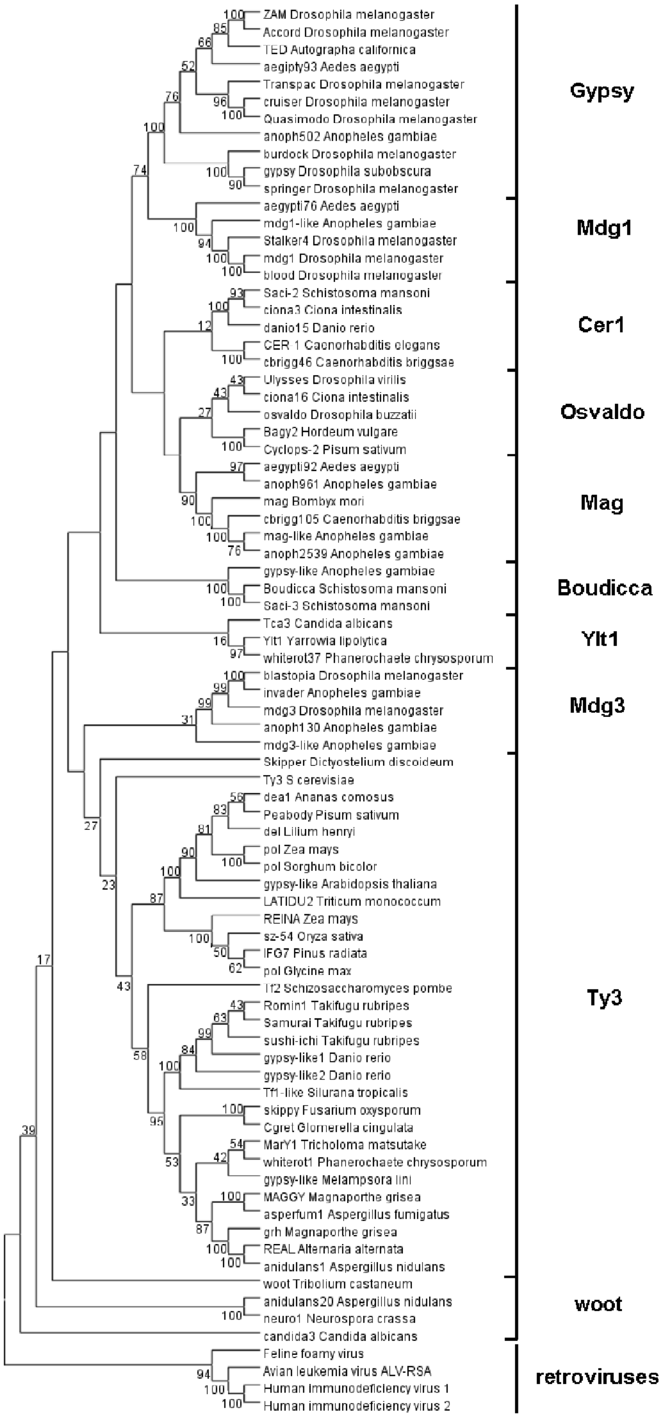

100
ZAM Drosophila melanogaster
85
Accord Drosophila melanogaster
66
TED Autographa californica
52
aegipty93 Aedes aegypti
76
Transpac Drosophila melanogaster
96
cruiser Drosophila melanogaster
100
Quasimodo Drosophila melanogaster
100
anoph502 Anopheles gambiae
burdock Drosophila melanogaster
74
100
gypsy Drosophila subobscura
90
springer Drosophila melanogaster
aegypti76 Aedes aegypti
100
mdg1-like Anopheles gambiae
94
Stalker4 Drosophila melanogaster
100
mdg1 Drosophila melanogaster
100
blood Drosophila melanogaster
93
Saci-2 Schistosoma mansoni
100
ciona3 Ciona intestinalis
12
danio15 Danio rerio
CER 1 Caenorhabditis elegans
100
cbrigg46 Caenorhabditis briggsae
43
Ulysses Drosophila virilis
43
ciona16 Ciona intestinalis
27
osvaldo Drosophila buzzatii
Bagy2 Hordeum vulgare
100
Cyclops-2 Pisum sativum
97
aegypti92 Aedes aegypti
anoph961 Anopheles gambiae
90
mag Bombyx mori
100
cbrigg105 Caenorhabditis briggsae
100
mag-like Anopheles gambiae
76
anoph2539 Anopheles gambiae
gypsy-like Anopheles gambiae
100
Boudicca Schistosoma mansoni
100
Saci-3 Schistosoma mansoni
Tca3 Candida albicans
16
Ylt1 Yarrowia lipolytica
97
whiterot37 Phanerochaete chrysosporum
100
blastopia Drosophila melanogaster
99
invader Anopheles gambiae
99
mdg3 Drosophila melanogaster
31
anoph130 Anopheles gambiae
mdg3-like Anopheles gambiae
Skipper Dictyostelium discoideum
Ty3 S cerevisiae
56
dea1 Ananas comosus
83
Peabody Pisum sativum
27
81
del Lilium henryi
90
pol Zea mays
100
100
pol Sorghum bicolor
gypsy-like Arabidopsis thaliana
23
87
LATIDU2 Triticum monococcum
REINA Zea mays
100
sz-54 Oryza sativa
50
IFG7 Pinus radiata
17
62
pol Glycine max
43
Tf2 Schizosaccharomyces pombe
43
Romin1 Takifugu rubripes
63
Samurai Takifugu rubripes
99
sushi-ichi Takifugu rubripes
84
58
gypsy-like1 Danio rerio
100
gypsy-like2 Danio rerio
Tf1-like Silurana tropicalis
100
skippy Fusarium oxysporum
95
Cgret Glomerella cingulata
39
54
MarY1 Tricholoma matsutake
42
53
whiterot1 Phanerochaete chrysosporum
gypsy-like Melampsora lini
100
MAGGY Magnaporthe grisea
33
asperfum1 Aspergillus fumigatus
87
grh Magnaporthe grisea
100
REAL Alternaria alternata
100
anidulans1 Aspergillus nidulans
woot Tribolium castaneum
anidulans20 Aspergillus nidulans
100
neuro1 Neurospora crassa
candida3 Candida albicans
Feline foamy virus
94
Avian leukemia virus ALV-RSA
100
Human immunodeficiency virus 1
100
Human immunodeficiency virus 2
Gypsy
Mdg1
Cer1
Osvaldo
Mag
Boudicca
Ylt1
Mdg3
Ty3
woot
retroviruses

Fig. Phylogenetic tree of Ty3/gypsy group based on the sum of amino acids in the RT, RNH and INT domains (approximately 550 amino acid positions).

References

Kumar S., Tamura K., Jakobsen I.B., Nei M. MEGA2: molecular evolutionary genetics analysis software // Bioinformatics. 2001. V. 17. P. 1244–1255.

Malik H.S., Eickbush T.H. Modular evolution of the integrase domain in the Ty3/Gypsy class of LTR retrotransposons // J. of Virol. 1999. V. 73. P. 5186–5190.

Malik H.S., Eickbuch T.H. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses // Genome Res. 2001. V. 11. P. 1187–1197.

Thompson J.D., Higgins D.G., Gibson T.J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. V. 22. P. 4673–4680.

Xiong Y., Eickbush T.H. Origin and evolution of retroelements based upon their reverse transcriptase sequences // The EMBO J. 1990. V. 9. P. 3353–3362.

EVOLUTION OF DIPLOID PROGENITORS OF COMMON WHEAT AS SUGGESTED BY ANALYSIS OF RAPD AND SUBTELOMERIC REPEATS

*Salina E.A.**[1], *Adonina I.G.*[1], *Lim Y.K.*[2], *Shcherban' A.B.*[1], *Vatolina T.Yu.*[1], *Leitch A.*[2]

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] Queen Mary University, London, UK
* Corresponding author: e-mail: salina@bionet.nsc.ru

Keywords: *evolution, RAPD, subtelomeric repeats, Aegilops*

Summary

Motivation: To study the correlation between species evolution and the divergence of repetitive sequences located at subtelomeric regions of the chromosomes.

Results: The *Aegilops* section Sitopsis is divided into two distinct groups: one containing only *Ae. speltoides* and the other containing *Ae. longissima*, *Ae. searsii*, *Ae. sharonensis* and *Ae. bicornis*. This grouping, by RAPD analysis, is in agreement with the taxonomical classification of the genus subsections. Here an evaluation of the divergence of repetitive sequences during evolution has been based on the primary sequence structure, the organization of the repetitive families, the copy numbers and chromosomal localization of two subtelomeric repetitive sequence families – Spelt1 and Spelt52. Patterns of repeat sequence amplification accompany *Sitopsis* species formation. However the divergence in primary structure of repeats and their subgenomic organization does not correlate with species divergence, but occurred over more ancient timescales.

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

Introduction

Five *Aegilops* species (all $n = 7$), namely *Ae. speltoides* Tausch., *Ae. longissima* Schw. & Mushc., *Ae. sharonensis* Eig., *Ae. bicornis* (Forssk) Jaub. & Sp., and *Ae. searsii* Feld.&Kis., form section *Sitopsis*. Of these, *Ae. speltoides* (genome SS) is the closest relative of the B and G genomes of polyploids wheats. *Aegilops speltoides* is the only outbreeding S-genome species and, as a consequence, is the most polymorphic. Repetitive sequences are the highly variable sequences in plant genomes. The variation is generated by sequence structure, organization and copy number and divergence in these characters may be influenced by species divergence. The aim of this study was to clarify phylogenetic relationship between species in Sitopsis using RAPD analysis and to evaluate the rate of repetitive sequence evolution in relation to species divergence in *Aegilops*. Two subtelomerically organized repetitive DNA sequence families were examined, with emphasis on sequence primary structure, genomic organization, copy numbers and chromosomal locations.

Materials and Methods

Plants DNAs: Fifteen lines of diploid species, *Ae. speltoides, Ae. longissima , Ae. sharonensis, Ae. bicornis,* and *Ae. Searsii* were used. Seeds of *Ae. speltoides* (lines TS01, TS05, TS41, TS42, and TS47), *Ae. longissima* (TL03, TL04, TL05, TL09 and TL05), *Ae. sharonensis* (TH01 and TH02), *Ae. bicornis* (TB05), and *Ae. searsii* (TE12, TE16) were obtained from Prof. M. Feldman (The Weizmann Institute of Science, Israel). As references, a single accessions each of *Triticum aestivum*, *Triticum durum* and *Triticum timopheevii* was also used.

DNA probes: Spelt1 and Spelt52 probes were as described previously (Salina *et al.*, 1998, Salina *et al.*, 2004).

Polymerase chain reaction (PCR) was as described previously (Salina *et al.*, 1998).

RAPD data analysis: RAPD patterns obtained using 18 primers were analyzed. Each PCR band was treated as a unit character and all species were scored for presence (1) or absence (0) of individual bands. The data matrix was used to calculate (dis) similarity coefficients following Nei and Li (1979) by PHYLIP software package (Felsenstein, 1989). The resulting distance matrices were used to construct a Neighbor-Joining phenogram using MEGA version 2.1 (Kumar *et al.*, 2001). The reliability of Neighbor-Joining tree was evaluated by bootstrap analysis (Felsenstein, 1989).

Dot-blot hybridization was carried out using standard procedures (Salina *et al.*, 1998). Repeat content was determined by comparing the hybridization signal intensities of genomic dilutions of studied lines and control dilutions.

In situ hybridization was as described previously (Salina *et al.*, 2004)

DNA cloning, sequencing and analysis was as described previously (Salina *et al.*, 1998).

Results and Discussion

RAPD analysis: Prelimary screening of 100 RAPDs identified 18 most polymorphic random primers for use in assessing *Aegilops* species relationships. Two hundred and eight bands were reproducibly obtained by RAPD amplification of 15 *Aegilops* lines. On the basis of presence or absence of individual bands, we constructed a dendrogram showing phylogenic relationships between *Sitopsis* species (Fig.). The section Sitopsis is divided into two distinct groups; one containing only *Ae. speltoides* and the other the remaining species, *Ae. longissima*, *Ae. searsii*, *Ae. sharonensis* and *Ae. bicornis*. This grouping by RAPD analysis is in agreement with the taxonomical classification of the subsections.

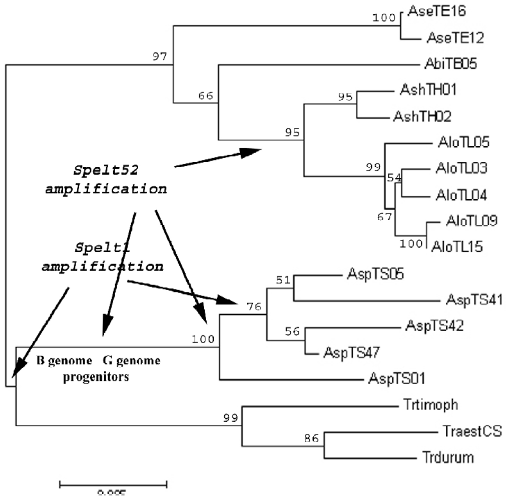


Fig. Dendrogram showing the genetic relationships for *Sitopsis* species.

Dot-blot and fluorescent in situ hybridization (FISH) of Spelt1 and Spelt52 with Aegilops species: The amount of Spelt1 and Spelt52 in fifteen accessions of *Aegilops* species was estimated by dot-blot hybridization. Using these methods the only species where Spelt1 repeats can be detected is in *Ae. speltoides* where there are usually about 10^5 copies per genome (lines, TS05, TS41, TS42, and TS47). The *Aegilops speltoides* line TS01 has a drastically decreased content of Spelt1 sequences compared with other lines and populations of this species. Spelt52 is present in *Ae. speltoides, Ae. longissima* and *Ae. sharonensis* but there is a large interspecific variation in the copy number of the repeat. Both repeats are absent in *Ae. bicornis* and *Ae. searsii.*

FISH analysis using Spelt1 and Spelt52 was carried out on *Aegilops* TS42, TS05, TS01, TH01, TL03. The work confirmed that both repeats localized to subtelomeric regions of the chromosomes. In line TS01, Spelt52 blocks occurred at subtelomeric regions to about half of the chromosome arms and Spelt1 occurred as a subtelomeric block on only one chromosome arm. In contrast, in other lines of *Ae. speltoides*, the signals of the Spelt1 and Spelt52 repeats occurred on 12–14 chromosome arms of the haploid genome. The total number of Spelt52 blocks per haploid genome was 5 for *Ae. longissima* line TL03, and 7 for *Ae. sharonensis* line TH01.

When these results are compared with RAPD dendrogram, it is apparent that the amplification of Spelt1 and Spelt52 repeats is accompanying the process of species divergence. Spelt1 sequences amplified several time during *Ae. speltoides* divergence. The amplification of Spelt52 occurs independently on two branches of section Sitopsis phylogenetic tree (Fig.).

Natural tetraploid and hexaploid wheat species have only Spelt1 repeats (Pestsova *et al*., 1998) with the exception of one accessions of *T. timopheevii* which also has Spelt52 (homologous to pGc1R-1) (Zang *et al*., 2002). On the base of these data we predict that the ancestor of B/G genome polyploids wheat was formed from an ancestor of *Ae. speltoides*. This agrees with the unique status of the S genome of *Ae. speltoides* as a putative donor of B and G genome polyploids wheats.

The sequencing of Spelt1 sequences from genomes of various Aegilops species: Sequence analysis of Spelt1 clones published previously (Salina *et al*., 1998) suggests that the Spelt1 repeat family is highly conserved in the genome of *Ae. speltoides*. We investigated whether sequence conservation was maintained between species in section Sitopsis. Using PCR, five *Aegilops* species were analyzed, *Ae. speltoides* line TS05*, Ae. longissima* line TL04*, Ae. sharonensis* line TH02*, Ae. bicornis* line TB05*,* and *Ae. searsii* line TE16. In addition we studied *Ae. speltoides* line TS01 with the lowest content of Spelt1 sequences. To clone the Spelt1 monomer, PCR of genomic DNA was performed with primers Spelt1L and Spelt1R, which are specific for this repeat. Sequence analysis of cloned amplicons with a length of 171 bp detected high (97–99 %) interspecific sequence identity.

Spelt1 showed high levels of sequence conservation and there was no indication that the sequence divergence correlated with speciation events as predicted from RAPD divergence patterns. There were clear indications of independent multiple amplification events in the genome of *Ae. speltoides*. Thus, the Spelt1 family copy number of 10^5 per genome in *Ae. speltoides* is much higher than in other *Aegilops* species in section *Sitopsis* where it is only detectable by PCR.

Organization Spelt52.1 and Spelt52.2 in genomes of various Aegilops species: Comparisons of cloned sequences the sequences from the GenBank database allowed two monomers types Spelt52.1 and Spelt52.2 to be identified. These are comprised of the common homologous Spelt52 sequence with a length of 277–280 bp and a region without homology of length 100 bp. To study the structural organization of the two types of monomers of Spelt52 repeat in the genomes of *Aegilops* section Sitopsis, we used PCR and primers (Sp52L, Sp52R, Ins1L, Ins1R, Ins2L, Ins2R) to homologous and non-homologous regions of Spelt52.1 and Spelt52.2. Genomic DNAs of each Aegilops species or line were used as templates. Several combinations of monomer types is possible: A-type - Spelt52.1 and Spelt52.2 monomers alternate with each other; B-type - duplication or tandem of Spelt52.1; C-type - duplication or tandem of Spelt52.2. Results of PCR show that of *Ae. speltoides* lines TS05, TS41, TS42, and TS47 are characterized by the A type repeat organization. For all primer combinations, PCR to line TS01 indicates the presence of all three possible combinations of repeats in the genome. The *Ae. longissima* (TL03, TL04, TL05, and TL09), *Ae. sharonensis* (TH01 and TH02), *Ae. bicornis* (TB05) have a similar and predominant alternation in A- and C-type of repeat. We failed to obtain amplicons using *Ae. searsii* (TE12 and TE16), presumably, due to a divergence in the region of primers. The presence of the Spelt52 repeat types is inconsistent with amplification events in *Aegilops* genome as predicted from RAPD divergence patterns and dot-blot hybridization (Fig.). There is no correlation between species divergence and changes to the repeat arrangement.

In summary, it should be noted that amplification of repeats is accompanying the speciation in *Aegilops* section *Sitopsis* but the divergence in primary sequence structure or the genomic organization of the repetitive families does not correlate with phylogenetic schemes generated by analysis of RAPD data.

Acknowledgements

The work was supported by INTAS (grant No. 01-0537), Russian Foundation for Basic Research (grant No. 02-04-48113), and Basic Research Program of RAN (No. 12.5).

References

Felsenstein J. PHYLIP: Phylogeny Inference Package (Version 3.2) // *Cladistics*. 1989. V. 5. P. 164–166.

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

Nei M., Li W.H. Mathematical model for studying genetic variation in terms of restriction endonucleases // Proc. Natl Acad. Sci. USA. 1979. V. 76. P. 5269–5273.

Pestsova E.G., Goncharov N.P., Salina E.A. Elimination of a tandem repeat of telomeric heterochromatin during evolution of wheat // Theor. Appl. Genet. 1998. V. 97. P. 1380–1386.

Salina E.A., Pestsova E.G., Adonina I.G., Vershinin A.V. Identification of a new family of tandem repeats in *Triticeae* genomes // Euphytica. 1998. V. 100. P. 231–237.

Salina E.A., Numerova O.M., Ozkan H., Feldman M. Alterations in subtelomeric tandem repeats during early stages of allopolyploidy in wheat // Genome. 2004, in press.

Zhang P., Friebe B., Gill B.S. Variation in the distribution of a genome-specific DNA sequences on chromosomes reveals evolutionary relations in the *Triticum* and *Aegilops* complex // Plant Syst. Evol. 2002. V. 235. P. 169–179.

COMPUTING LARGE PHYLOGENIES WITH STATISTICAL METHODS: PROBLEMS AND SOLUTIONS

Stamatakis A.P.*, Ludwig T., Meier H.

Department of Computer Science, Technische Universität München; Department of Computer Science, Ruprecht-Karls Universität Heidelberg
* Corresponding author: e-mail: stamatak@in.tum.de

Keywords: *evolution, phylogenetics, maximum likelihood, large phylogenies*

Summary

The computation of ever larger as well as more accurate phylogenetic trees with the ultimate goal to compute the “tree of life” represents a major challenge in Bioinformatics. Statistical methods for phylogenetic analysis such as maximum likelihood or bayesian inference, have shown to be the most accurate methods for tree reconstruction. Unfortunately, the size of trees which can be computed in reasonable time is limited by the severe computational complexity induced by these statistical methods.

However, the field has witnessed great algorithmic advances over the last 3 years which enable inference of large phylogenetic trees containing 500–1000 sequences on a single CPU within a couple of hours using maximum likelihood programs such as RAxML and PHYML. An additional order of magnitude in terms of computable tree sizes can be obtained by parallelizing these new programs.

In this paper we briefly present the MPI-based parallel implementation of RAxML (Randomized Axelerated Maximum Likelihood), as a solution to compute large phylogenies. Within this context, we describe how parallel RAxML has been used to compute the – to the best of our knowledge-first maximum likelihood-based phylogenetic tree containing 10.000 taxa on an inexpensive LINUX PC-Cluster.

In addition, we address unresolved problems, which arise when computing large phylogenies for real-world sequence data consisting of more than 1.000 organisms with maximum likelihood, based on our experience with RAxML. Finally, we discuss potential algorithmic and technical enhancements of RAxML within the context of future work.

Availability: wwwbode.in.tum.de/~stamatak.

Introduction

The inference of large phylogenetic trees based upon statistical models of nucleotide substitution is computationally intensive since the number of potential alternative tree topologies grows exponentially with the number of sequences and due to the high computational cost of the likelihood evaluation function for each individual topology. Although this has not been demonstrated to date, it is widely believed that maximum likelihood-based phylogenetic analysis is an NP-complete problem.

Therefore, progress in this field, in terms of gain in several orders of magnitude in conjunction with inexpensive hardware requirements, is rather achieved by algorithmic optimizations and introduction of new heuristics than by brute-force allocation of all available computational resources. E.g. a large and expensive grid of supercomputers has been used to conduct one of the most computationally intensive phylogenetic analyses to date based on the relatively slow and old parallel fastDNAml (Stewart *et al.*, 2001) code within the framework of the HPC challenge at the 2003 Supercomputing Conference (for details see www.sc-conference.org/sc2003/tech_hpc.php). Despite the unchallenged technical success the extreme computational effort could have been avoided by using more recent algorithms which execute approximately 50 times faster than fastDNAml and yield better results at the same time.

In a survey conducted by T.Williams *et al.* (2003) its has been demonstrated that MrBayes (Huelsenbeck *et al.*, 2001a), an implementation of bayesian phylogenetic inference based on the Metropolis-Coupled Markov Chain Monte Carlo technique, appears to be the currently fastest and most accurate program for phylogenetic inference. However, this survey is based entirely on simulated data, which can potentially generate misleading results.

More recently, Guidon *et al.* (2003) released a program called PHYML which is equally accurate and significantly faster than MrBayes and some of the most popular or efficient maximum likelihood programs like MetaPIGA (Lemmon *et al.*, 2002), PAUP (Swofford *et al.*, 2004), treepuzzle (Schmidt *et al.*, 2001) and fastDNAml (Olsen *et al.*, 1994).

Thus, PHYML and MrBayes represent the -to the best of our knowledge- currently fastest and most accurate phylogeny programs. In a recent paper (Stamatakis *et al.*, 2004a) we describe the basic sequential implementation of RAxML (Randomized Axelerated Maximum Likelihood), which clearly outperforms PHYML and MrBayes on 9 large real world alignments containing 101 up to 1000 sequences both in terms of execution speed and final likelihood values, whereas it performs slightly worse on simulated data. Furthermore, in (Stamatakis *et al.*, 2004) we also show that MrBayes fails to converge or converges significantly slower than RAxML and PHYML within reasonable time limits for some real world data sets. This result is not an argument against bayesian methods which are very useful and have experienced great impact (Huelsenbeck *et al.*, 2001b) but for maximum likelihood methods which are still significantly faster and useful for verifying results of bayesian analyses.

Parallelization

The basic sequential algorithm of RAxML is outlined in (Stamatakis *et al.*, 2004a). For parallelization we have chosen a coarse-grained approach which intends to minimze communication in order to allow for a http-based distributed implementation of RAxML as well (Stamatakis *et al.*, 2004b).

The topology optimization process of RAxML is based upon a fast pre-evaluation of a large number of alternative topologies by application of the subtree rearrangement technique, which is also known as subtree pruning & re-grafting. The parallel code is based on a simple master-worker architecture, where the master maintains the currently best tree and distributes work by subtree IDs which are represented by simple integer values. Each subtree is then individually rearranged within the currently best tree by a worker. When a rearrangement step has been completed, i.e. all subtrees of the current tree have been rearranged, the best 20 (or # of workers, whichever is higher) trees obtained from this step are gathered by the master. The master then redistributes those 20 trees to the workers for branch length optimization and commences a new cycle of subtree rearrangements with the updated best tree. This process is repeated until no better tree is found.

However, the sequential algorithm contains a closely-coupled step: the subsequent application of topological improvements (Stamatakis *et al.*, 2004a) which is difficult to parallelize. Thus, we have chosen to introduce some non-determinism in the parallel program to solve this problem. The non-determinism in the parallel program leads to a traversal of tree space on different paths for each individual program execution. As demonstrated by experimental results this non-determinism does not impose serious restrictions on program performance and partially leads to even superlinear speedup values. In Figure we plot the average speedup values for a 1.000 taxon alignment which has been extracted form the ARB small subunit ribonucleic acid database (Ludwig *et al.*, 2004) over 4 parallel RAxML runs on 4, 8, 16, and 32 2.66 GHz Xeon processors respectively. Due to the non-determinism of the parallel code we provide two types of speedup values: "Fair" speedup indicates the point of time at which the parallel code detects a tree which shows a better likelihood value than the final topology of the sequential execution and "normal" speedup indicates the standard definition accounting for execution time until termination.

Computation of a 10.000-taxon phylogeny with RAxML

In order to conduct a large and meaningful phylogenetic analysis with RAxML we extracted an alignment comprising 10.000 sequences including organisms of the three domains Eukarya, Bacteria, and Archaea from the ARB database. The computation of the 10.000-taxon tree was conducted using the sequential, as well as the parallel version of RAxML. One of the advantages of RAxML consists in the randomized generation of parsimony starting trees. Thus, we computed 5 distinct randomized parsimony starting trees sequentially along with the first 3–4 rearrangement steps on a small cluster of Intel Xeon 2.4GHz processors at our institute. This phase required an average of 112.31 CPU hours per tree.

Thereafter, we executed several subsequent parallel runs (due to job run-time limitations of 24 hrs) starting with the sequential trees on either 32 or 64 processors on the 2.66GHz cluster mentioned above. The parallel computation required an average of approximately 1.600 accumulated CPU hours per tree. The best likelihood obtained for the 10.000 taxa was -949570.16 the worst -950047.78 and the average -949867.27.

PHYML reached a likelihood value of -959514.50 after 117.25 hrs on a 64-bit Itanium2 processor. Note, that the parsimony starting trees computed with RAxML showed likelihood values ranging between -954579.75 and -955308.00. The average time required for computing those starting trees on the Xeon processor was 10.99 hrs. Since bootstrapping is not feasible for this large data size and in order to gain some basic information about similarities among the 5 final trees we built a majority-rule consensus tree with consense (Jermiin *et al.*, 1997). The consensus tree has 4777 bifurcating inner nodes which appear in all 5 trees, 1046 in 4, 1394 in 3, 1323 in 2, and 1153 in only 1 tree (average: 3.72). The results from this large phylogenetic analysis including all final trees as well as the consensus tree are available at: wwwbode.cs.tum.edu/~stamatak.

The final version of this paper will also include a biological analysis of the 10.000-taxon phylogeny.

Problems

Several new problems arise within the context of computation of large trees. An important observation is that memory consumption becomes critical, e.g. MrBayes and PHYML fail to execute for the 10.000-taxon alignment on a 32 bit processor with 4MB of main memory due to excessive memory requirements. Moreover, MrBayes could not be ported to a 64 bit Itanium2 processor whereas PHYML finally required 8.8MB of memory. In contrast to MrBayes and PHYML, RAxML required only approximately 800MB for the 10.000-taxon alignment. Thus, phylogeny programs for computation of large trees need to be designed for low memory consumption, since 64 bit architectures also induce a significant additional cost factor. Furthermore, consense is apparently not able to handle more than 5 10.000-taxon trees since it constantly exited with an error message when executed with more than 5 input trees.

Another important problem which is often underestimated is tree visualization which requires novel concepts for displaying large trees. Information obtained by phylogenetic analysis becomes valuable and can be interpreted only if appropriate tools are available. An initial visualization of the 10.000-taxon phylogeny with ATV (Zmasek *et al.*, 2001) demonstrated that this standard tool is completely inadequate for viewing large trees. In fact, 2-D and 3-D hyperbolic tree viewers such as Walrus or Hypertree have been proposed as a solution (for details see www.caida.org/tools/vizualisation/walrus and hypertree.sourceforge.net) for large trees and graphs, which we did not find very helpful in the specific case however.

Finally, the assignment of confidence values to large trees remains problematic since execution of typically 100 or 1.000 distinct inferences to obtain bootstrap values or build consensus trees does not appear to be computationally feasible at present. In addition, MrBayes which directly yields confidence values is presently too slow and requires an excessive amount of memory for this tree sizes.

Thus, apart from the necessary improvements of associated tools phylogeny programs still require to become faster and yield at least equally good trees at the same time. In addition, they should incorporate more complex and exact models of evolution. Those two basic directions of research represent controversial targets due to an apparent trade-off between speed and quality. More sophisticated models, such as for example the General Time Reversible Model (GTR) of nucleotide substitution compared to HKY85 lead to significantly increased execution times.

Conclusion, Current and Future Work

Along with PHYML, RAxML currently represents one of the fastest and most accurate sequential phylogeny programs. In contrast to PHYML which is unfortunately only available as sequential program, we also provide a parallel MPI-based implementation of RAxML which has been used to conduct the – to the best of our knowledge – largest maximum likelihood analysis to date on a medium-sized PC cluster. Our program along with a benchmark set of best-known trees for real-world alignments, which have all been obtained by RAxML, is freely available at wwwbode.in.tum.de/~stamatak.

Currently, we are implementing model parameter optimization for the HKY85 and GTR models of nucleotide substitution in the new sequential version of RAxML, which will soon be released. Furthermore, we are working on a RAxML-based tool for splitting-up alignments into overlapping sub-alignments within the context of a divide-and-conquer supertree approach to phylogenetic inference.

Future work will cover the exploitation of the intrinsic fine-grained parallelism of RAxML on likelihood vector level by using Graphics Processor Units (GPUs) or other inexpensive hardware in a similar way as introduced by Kruger *et al.* (2003) for numerical simulations.

Finally, we will analyze the effect of the application of divide-and-conquer approaches and associated supertree methods to large maximum likelihood analyses in terms of final tree quality and execution times.

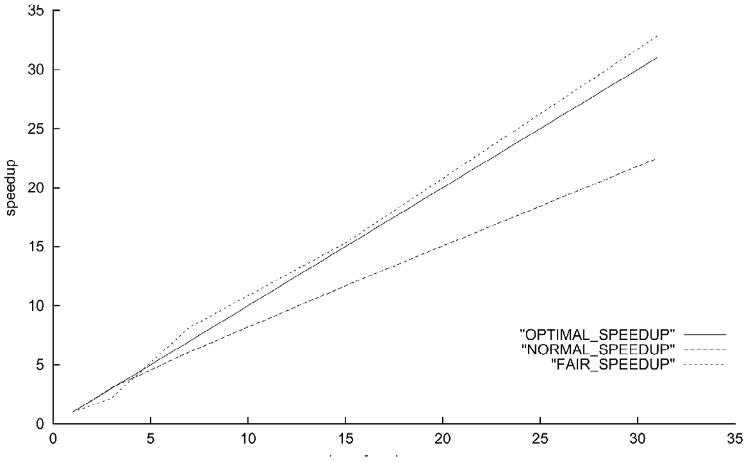


Fig. Normal and fair speedup values of parallel RAxML for a 1.000-taxon alignment on 4, 8, 16, and 32 Intel Xeon 2.66 GHz processors.

Acknowledgements

This work is sponsored under the project ID **ParBaum**, within the framework of the “Competence Network for Technical, Scientific High Performance Computing in Bavaria” (KONWIHR).

References

Guidon S., Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood // Syst. Biol. 2003. V. 52(5). P. 696–704.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. Bayesian Inference and its Impact on Evolutionary Biology // Science. 2001a. V. 294. P. 2310–2314.

Huelsenbeck J.P., Ronquist F. MrBayes: Bayesian Inference of Phylogenetic Trees // Bioinformatics. 2001b. V. 17(8). P. 754–755.

Jermiin L.S., Olsen G.J., Mengersen K.L. Easteal S. Majority-rule Consensus of Phylogenetic Trees Obtained by Maximum-Likelihood analysis // Mol. Biol. Evol. 1997. V. 14. P. 1297–1302.

Krüger J., Westermann R. Linear Algebra Operators for GPU Implementations of Numerical Algorithms. Proc. of SIGGRAPH2003. 2003.

Lemmon A., Milinkovitch M. The Metapopulation Genetic Algorithm: An Efficient Solution for the Problem of Large Phylogeny Estimation // Proc. Natl Acad. Sci. 2002. V. 99. P. 10516–10521.

Ludwig W. *et al.* ARB : A Software Environment for Sequence Data // Nucl. Acids Res. (2004. V. 32(4). P. 1363–1371.

Olsen G., Matsuda H., Hagstrom R., Overbeek R. FastDNAml: A Tool for Construction of Phylogenetic trees of DNA sequences using Maximum Likelihood // Comput. Applic. Biosci. 1994. V. 10. P. 41–48.

PAUP project site: paup.csit.fsu.edu, visited March 2004.

Schmidt H.A. *et al.* TREE-PUZZLE: Maximum Likelihood Pylogenetic Analysis using Quartets and Parallel Computing // Bioinformatics. 2002. V. 18. P. 502–504.

Stamatakis A.P., Ludwig T., Meier H. New Fast and Accurate Heuristics for Inference of Large Phylogenetic Trees. To be published in Proc. of IPDPS2004. 2004a. Preprint available on-line at: wwwbode.in.tum.de/~stamatak/publications.html

Stamatakis A.P., Ott M., Ludwig T., Meier H. DRAxML@home: A Distributed Program for Computation of Large Phylogenetic Trees. To be published in FGCS. 2004b.

Stewart C. *et al.* Parallel Implementation and Performance of fastDNAml – a Program for Maximum Likelihood Phylogenetic Inference // Proc. of SC2001. 2001.

Williams T.L, Moret B.M.E. An Investigation of Phylogenetic Likelihood Methods // Proc. of BIBE2003. 2003.

Zmasek C.M., Eddy M.R. ATV: Display and Manipulation of Annotated Phylogenetic Trees // Bioinformatics. 2001. V. 17(4). P. 383–384.

COUNTERING COORDINATED AMINO ACID SUBSTITUTIONS IN PHYLOGENETIC ANALYSIS

Triboy T.I., Sherbakov D.Yu.*

Limnological Institute SB RAS, Irkutsk, Russia
* Corresponding author: e-mail: sherb@lin.irk.ru

Keywords: *evolution, homoplasy, coordinated substitutions, phylogenetic analysis*

Summary

Motivation: Countering difficulties in phylogenetic inferences from amino acid sequences which due to homoplasies arising from coordinated evolution in some positions.

Results: It is shown that representing of the groups of amino acids which evolve in coordinated fashion by a single residue belonging to the group increases the precision of phylogenetic analysis.

Introduction

Phylogenetic analysis critically depends on compliance of the data to several conditions shared by the most if not all contemporary methods (i.e. [2]). One of such requirements is the independence of traits. Like the other assumptions of the analysis, it is not necessarily true. It is well known, that in amino acid sequences many substitutions are constrained by physico-chemical properties of the residues. These coordinated substitutions result in high correlations of certain properties such as amino acid volume, hydropathy or pK among positions scattered widely along the sequence [1]. The constrains are believed to ensure stability of the protein spatial structure and its functions.

From the viewpoint of phylogenetic inferences the coordinated amino acid substitutions are one of the most important sources of homoplasies [5, 6] which are likely to impede the results of analysis. If this is not taken into account, the risk of obtaining wrong tree topology may increase dramatically, which may result in wrong understanding of evolutionary processes.

Here we propose the approach allowing one to take into account coordinated substitutions in order to counter possible concerted homoplasy which is similar to method described by Hillis *et al.* [4] for morphological characters causing homoplasies. We use computer simulations to show that taking into account coordinated amino acid substitutions decreases the risk of wrong phylogenetic inferences in cases when amino acid substitutions in some sequence positions are robustly constrained by their physico-chemical properties.

Methods

Phylogenetic trees of random topology were obtained with program phylogen under assumption of constant birth/death ratio of evolutionary lineages. The tree length was adjusted so that 70–80 % of positions were variable.

The trees were used as the template along which a random protein sequence was evolved with program pseq-gen [7].

Tree topologies were inferred from the data sets consisting of variable number of amino acids of variable length with maximum likelihood method implemented in program PHYML [3].

Tree topologies were compared with program treedist from PHYLIP 3.6b package using symmetric distances.

Coordinated evolution was simulated by adding ten amino acid residues to the C-terminus of model protein sequences following the rules: 1. The data set was searched for ten first variable positions where pK of side residue would vary; 2. Depending on the amino acid in that position another one was added to the sequence according to the Table:

Template Amino acid	Amino acid added
D	K
E	R
K	D
R	E
Any other	A

Results and Discussion

We studied the impact of coordination of physic-chemical properties of amino acids on the precision of phylogenetic inferences by simulating evolution of random amino acid sequences along the random tree. The simulated date sets were used to infer new tree which was then compared quantitatively to the original one. In the context of this study the sequence set thus obtained was considered to be free of positions evolving in coordinated fashion. They may occur only occasionally depending on the model of molecular evolution used. We treated this case as the model for a sequence set from which all positions where only coordinated substitutions may occur are "masked" so that each group is represented by a single position.

Sequence length was varied from 100 to 1000 amino acids, number of sequences in the alignment varied from 20 to 100. Since the length of the sequence at the same degree of diversity does not influence the performance of tree inference noticeably (Fig. A), the coordinated evolution was simulated by adding ten amino acids to the C end of the sequences as described in "Methods".

For each combination of parameters simulation was run ten times. The results are presented in Fig. The tree topology, obtained for a sequence set containing amino acid stretch modeling coordinated evolution differs markedly from the original tree. As expected, the difference is most dramatic for large number of short sequences.

Our unpublished data on the examination of real data sets consisting of sufficiently far diverged amino acid sequences suggest that normally about 10 % of variable positions contain amino acids with highly correlated certain chemical properties. The ratio of such positions depends considerably of the protein analyzed. In model datasets where no special efforts were made to obtain this correlation, the ratio is much less (data not shown).

Here we show that in cases when the ratio of correlated properties of amino acids is high and therefore the constrains on the substitutions is high, in order to obtain reliable phylogeny one should filter for groups of correlated positions representing each of them by a single position and only then start the phylogenetic inference.

Application of this approach to several alignments of amino acid sequences of the Folmer fragment of mitochondrial gene coding for the first sub unit of cytochrome oxydase allowed us to obtain trees which corroborate well both with bio-geographic and morphological data on respective species while traditional approach resulted in paradoxical results.

Acknowledgements

This work was supported by RFBR grant No. 04-04-48949, Russian/Austrian grant RFBR No. 03-04-20019 BNST, Russian/Dutch grant RFBR No. 04-01-89000 NWO, NOW No. 047.016.013. Authors are grateful to S. Semovsky, D. Afonnikov and E. Verheyen for helpful discussions and suggestions.

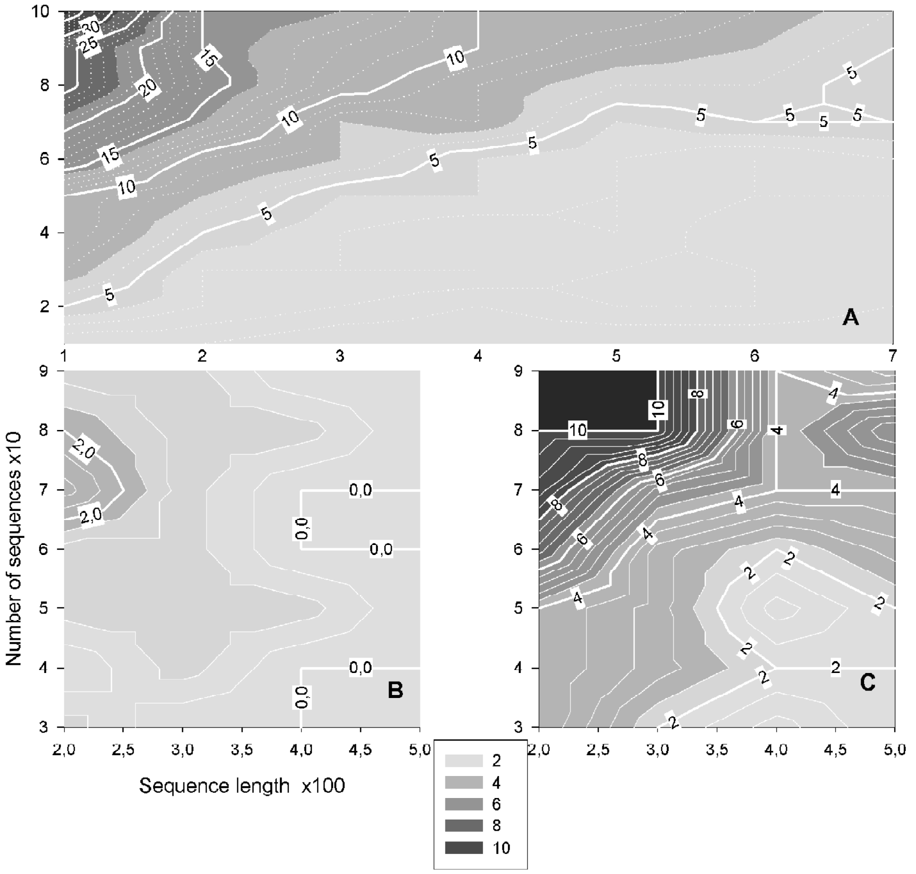


Fig. Difference between model and inferred trees. A – performance of the tree inference over wide range of parameters; B – coordinated positions are "masked", C – sequences contain amino acid stretches modeling coordinated evolution.

References

1. Afonnikov D.A., Oshchepkov D.Y., Kolchanov N.A. Detection of conserved physico-chemical characteristics of proteins by analyzing clusters of positions with co-ordinated substitutions // Bioinformatics, 2001. V. 17. P. 1035–1046.
2. Davis J.I., Simmons M.P., Stevenson D.W., Wendel J.F. Data decisiveness, data quality, and incongruence in phylogenetic analysis: an example from the monocotyledons using mitochondrial atp A sequences // Syst. Biol. 1998. V. 47. P. 282–310.
3. Guindon S., Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood // Syst. Biol. 2003. V. 52. P. 696–704.
4. Harris S.R., Wilkinson M., Harques C. Countering concerted homoplasy // Cladistics. 2003. V. 19. P. 128–130.
5. Hassanin A., Lecointre G., Tillier S. The 'evolutionary signal' of homoplasy in protein-coding gene sequences and its consequences for a priori weighting in phylogeny // C.R. Acad. Sci. III. 1998. V. 321. P. 611–620.
6. Haszprunar G. Parsimony analysis as a specific kind of homology estimation and the implications for character weighting // Mol. Phylogenet. Evol. 1998. V. 9. P. 333–339.
7. Rambaut A., Grassly N.C. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees // Comput. Appl. Biosci. 1997. V. 13. P. 235–238.

DOMINANT MODE IN *PISUM* DIVERSITY GENERATION: WHAT IS THE IMPACT OF TRANSPOSABLE ELEMENTS?

Vershinin A.V.[1,2], **Allnutt T.R.**[2], **Knox M.R.**[2], **Ambrose M.J.**[2], **Ellis T.H.N.**[2]*

[1] Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia; [2] John Innes Centre, Norwich, UK
* Corresponding author: e-mail: avershin@bionet.nsc.ru

Keywords: *transposable elements, Pisum evolution, molecular markers, recombination, domestication, retrotransposon activity*

Summary

We studied the evolutionary history of the genus *Pisum*, using SSAP markers. They allowed to assess insertion site polymorphism of a representative of each of the two major groups of LTR-containing retrotransposons, *PDR1* (Ty1/*copia*-like) and *Cyclops* (Ty3/*gypsy*-like), together with *Pis1*, a member of the *En/Spm* transposon superfamily. These markers showed that *Pisum* is exceptionally polymorphic for an inbreeding species. The patterns of phylogenetic relationships deduced from these three families were in general agreement. The retrotransposon derived markers gave a clearer separation of the main lineages than the *Pis1* markers. There were more species-specific and unique *PDR1* markers in *P. fulvum* and *P. elatius* compared to *Pis1* markers, pointing to *PDR1* activity during speciation and diversification, but the proportion of these markers is low. The overall genetic diversity of *Pisum* and the extreme polymorphism in all species, except *P. abyssinicum*, indicate a high contribution of recombination between multiple ancestral lineages compared to transposition within lineages. The two independently domesticated pea species, *P. abyssinicum* and *P. sativum*, arose in contrasting ways from the common processes of hybridisation, introgression and selection without associated transpositional activity.

Introduction

The molecular mechanisms that generate genetic diversity are bewilderingly complex. The advent of genome sequencing have facilitated genome comparison; however, these approaches are very expensive and time consuming. The use of genetic markers is an alternative and less costly surveying method. Mobile genetic elements (or transposable elements, TEs) are a major component of plant genomes where they may comprise more than 50 % of the nuclear DNA (SanMiguel, Bennetzen, 1998) and contribute to diversity through both insertion site polymorphism and small structural rearrangements (Bennetzen, 2000). Transposable elements are classified into two groups according to their transposition mechanism and mode of propagation (Finnegan, 1992): retrotransposons transpose via an RNA intermediate, whereas transposons move by excision and reintegration (“cut-paste”). The ubiquity and distribution of these TEs suggest that they should be potentially useful as diagnostic tools conforming to the requirement for abundant and reliable markers of biodiversity.

Our approach to study the genetic structure and evolutionary history of genus *Pisum* is based on an SSAP, sequence-specific amplification polymorphism, which is a multiplex AFLP-like gel-based method (Ellis *et al.*, 1998). Specific sequences are different classes of transposable elements (TEs): LTR-containing retrotransposons, *PDR1*, the representative of Ty1/*copia* group, and *Cyclops* belonging to the Ty3/*gypsy* group (Chavanne *et al.*, 1998), together with *Pis1*, a member of the *En/Spm* transposon superfamily (Shirsat, 1988). We designed specific primers to 3’-end of these sequences to generate markers derived from genomic sequences neighbouring retrotransposon ends.

Methodological Approach

Analysis of germplasm collection. We analysed the John Innes Centre pea germplasm collection, which contains about 3000 accessions. Such size of collection represents a great challenge because until now there were no routine methods of polymorphism detection and there were no computer programmes adapted to so large number of accessions. Therefore, we sub-divided the analysis on three steps: 1). Preliminary survey of 56 accessions selected on geographical and morphological criteria gave us a **reference set** of accessions; 2). The analysis of **extended sets** of the 4 main *Pisum* species together with the reference set; 3). A **final set** included 52 representatives of the main branches from phylogenetic trees constructed for the extended sets. It consisted of the 10 accessions of *P. fulvum*, 12 accessions of *P. elatius*, 5 accessions of *P. abyssinicum*, 2 accessions of *P. humile* and 25 accessions of *P. sativum*.

The total number of pair-wise mis-matches was used to calculate pair-wise distances from the SSAP band scores. The resulting matrix was used as input for principal component analysis (PCA) using the GenStat 5 package (http://www.nag.co.uk/stats/tt_soft.asp) and analysis of molecular variance (AMOVA) (Excoffier, Smouse, and Quattro 1992). Phylogenetic trees were generated from this matrix using the Neighbour-Joining (NJ) algorithm of the NEIGHBOR program in the PHYLIP package (Felsenstein, 1993). NJ dendrograms were also constructed from pair-wise Φ_{ST} values between accession groups generated by AMOVA. Trees were plotted using TREEVIEW (http://taxonomy.zoology.gla.ac.uk/).

Results

Do different TE classes show the same pattern of phylogenetic diversity? The markers derived from both retrotransposons showed very similar patterns of interspecies relationships with clear separation of the main lineages. The *Pis1*-derived markers gave similar results, but the discrimination was not as clear. The analysis of extended sets of the 4 main *Pisum* species revealed a distinct pattern of NJ tree for each basic lineage, which reflects the different evolutionary history of each species. The main characteristics of the combined tree for final set of accessions reiterated the conclusions from the analyses of the four groups independently and are: 1) Clear separation of the *P. fulvum* lineage; 2) Extremely low diversity within well distinct group of *P. abyssinicum*; 3) A broad distribution of *P.elatius* branches, some of which are intermingled with other lineages.

Evolutionary dynamics of genome microstructure. SSAP markers can be used to measure fluctuations in copy number between accessions and the appearance of unique insertions. We identified unique bands (that are present in only one accession), and species-specific bands that are present in accessions of only one species, regardless of their frequency. All species, except *P. abyssinicum*, showed an extremely high percentage of polymorphic bands (Table). However, while the percentage of species-specific (including unique) markers makes up a noticeable part of overall polymorphism of *P. fulvum* and *P. elatius*, within *P. sativum* the proportion of such markers is close to zero and similar to that of *P. abyssinicum* where none markers were unique. *PDR1* showed more than a two-fold higher percentage of species-specific and unique markers compared to *Pis1*, suggesting that its retrotransposition might have occurred during speciation of *P. fulvum* and *P. elatius*. However, the proportion of these markers is small and the element copy number is constant for all accessions. This is consistent with the introgression, segregation and small rearrangements, rather than transposition itself being dominant mode of diversity generation, even for the most ancient *Pisum* lineages, *P. fulvum* and *P. elatius*.

Genetic diversity associated with domestication. The accessions studied have a very similar *PDR1* copy number, well suited to SSAP analysis and the insertion sites are distributed throughout pea genome suggesting these are comprehensive diagnostic markers. We scored the number of markers shared by two, three and all species. *P. fulvum* has more markers in common with

P. abyssinicum than *P. elatius* (Fig. A) whereas the latter has the highest proportion of markers in common with *P.sativum*. Furthermore, *P. abyssinicum* and *P. sativum* do not have shared markers that are absent from the other lineages, supporting the idea that these lineages are of recent origin and were brought into cultivation independently. Close to zero species-specific and unique markers (Table) leads to the conclusion that *P. abyssinicum* and *P. sativum* arose in contrasting ways from the common processes of hybridization, introgression and selection without associated transpositional activity.

Table. Characteristics of markers derived from different TEs

Species	Transposable element	Total number of markers per species	Number of polymorphic markers within species	Number of speciesspecific markers	
				total	unique
P. fulvum	*PDR1* (259)*	167 (64.5 %)	160 (95.8 %)	35 (21.0 %)	17 (10.2 %)
	Cyclops(343)*	275 (80.2 %)	266 (96.7 %)	20 (7.3 %)	3 (1.1 %)
	Pis1 (215)*	159 (73.9 %)	157 (98.7 %)	16 (10.1 %)	6 (3.8 %)
P. elatius	*PDR1*	189 (73.0 %)	183 (96.8 %)	25 (13.2 %)	16 (8.5 %)
	Cyclops	312 (91.0 %)	304 (97.4 %)	7 (2.2 %)	2 (0.6 %)
	Pis1	158 (73.5%)	158 (100%)	12 (7.6 %)	6 (3.8 %)
P. abyssinicum	*PDR1*	75 (28.9%)	16 (21.3%)	2 (2.7 %)	0
	Cyclops	122 (35.5%)	5 (4.1%)	0	0
	Pis1	35 (16.3%)	5 (14.3%)	0	0
P. sativum	*PDR1*	159 (61.4%)	149 (93.7%)	4 (2.5 %)	3 (1.9 %)
	Cyclops	259 (75.5%)	240 (92.7%)	0	0
	Pis1	133 (61.9%)	131 (98.5%)	5 (3.7 %)	2 (1.5 %)

* Total number of markers identified in all species.

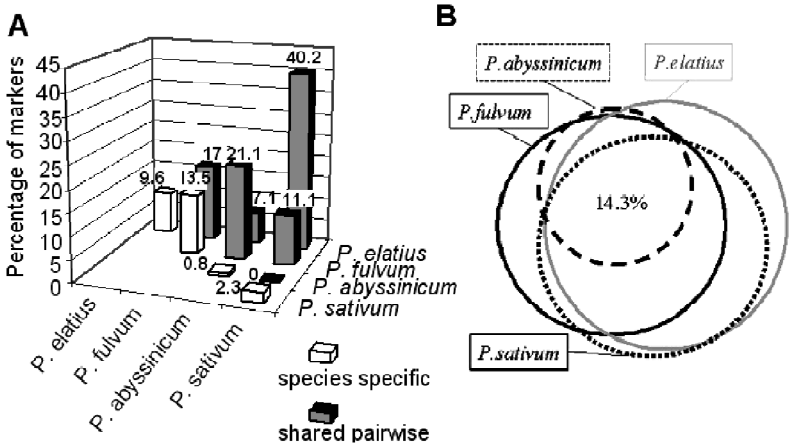


Fig. Evolutionary dynamics of different *Pisum* species demonstrated by the proportions of *PDR1* SSAP markers. **A**. Markers shared exclusively by species pairs. Diagonal columns show the proportions of species-specific markers. **B**. Circles illustrating the proportions of markers occurring in different species combinations.

***Pisum* as a species complex.** The general pattern of markers sharing is illustrated on Figure B where the overlapping circles depict the proportions of the corresponding gene pool of *Pisum*. The different segments reflect the proportions of species-specific and shared markers. The proportion of markers in common to all *Pisum* is 14 %, but these markers are not fixed markers and its polymorphism is shared between lineages. Thus, the remaining 86 % of the *Pisum* genome underestimates *Pisum* diversity. These results strongly suggest that *Pisum* is a species complex.

Conclusion

1. SSAP markers showed that *Pisum* is exceptionally polymorphic for an inbreeding species.

2. The patterns of phylogenetic relationships deduced from different TEs are in general agreement.

3. Introgression, segregation and small rearrangements, rather than transposition itself being the dominant mode of diversity generation, even for the most ancient *Pisum* lineages, *P. fulvum* and *P. elatius*.

4. Two independently domesticated species, *P. abyssinicum* and *P. sativum*, arose in contrasting ways from the common processes of hybridisation, introgression and selection without associated transpositional activity.

Acknowledgement

This work was supported by the EU project TEGERM QLK-CT-2000-01502.

References

Bennetzen J.L. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions // Plant Cell. 2000. V. 12. P. 1021–1029.

Chavanne F., Zhang D.-X., Liaud M.-F., Cerff R. Structure and evolution of Cyclops: a novel giant retrotransposon of the T*y3/Gypsy* family highly amplified in pea and other legume species // Plant Mol. Biol. 1998. V. 37. P. 363–375.

Ellis T.H.N., Poyser S.J., Knox M.R., Vershinin A.V., Ambrose M.J. Polymorphism of insertion sites of *Ty1*-copia class retrotransposons and its use for linkage and diversity analysis in pea // Mol. Gen. Genet. 1998. V. 260. P. 9–19.

EVOLUTION OF INTERLEUKIN-18 BINDING PROTEIN

Watanabe M.[1]***, Goto N.***[2]***, Watanabe Y.***[1]***, Nishiguchi S.***[1]***, Shimada K.***[1]***,***

Yasunaga T.[2]***, Yamanishi H.****[1]

[1] Hirakata Ryoikuen, Tsudahigashi 2-1-1, Hirakata, Osaka 573-0122, Japan; [2] Department of Human Genome Research, Genome Information Research Center,
Osaka University, Suita, Osaka 565-0871, Japan
* Corresponding author: e-mail: hirochan@hirakataryoiku-med.or.jp

Keywords: *IL-18, IL-18BP, IL-18BD, vertebrates, chordopoxviruses, computer analysis*

Summary

Motivation: Interleukin-18 (IL-18) is one of the pivotal cytokines, which controls defense mechanism called inflammation. To develop proteins available for controlling the local and/or circulating IL-18 levels, we initiated study of interleukin-18 binding proteins (IL-18BPs).

Results: Eighteen IL-18BPs, 7 from vertebrates and 11 from chordopoxviruses, were picked from the NCBI database. All of their IL-18 binding domains (IL-18BDs) were aligned and a phylogenetic tree constructed. Our results suggested that at least two independent events created two different ancestral viral IL-18BP genes by retroposing IL-18BP genes from the vertebrate lineage. These two events are estimated to have occurred, after an ancient mammalian IL-18BP gene diverged from bird, and before the mammalian IL-18BP gene diverges into human, ungulate and rodent IL-18BP genes.

Introduction

Interleukin-18 (IL-18) is a proinflammatory cytokine that plays a key role in the immune responses called inflammation. When inflammation goes awry, it can lead to heart attacks, rheumatoid arthritis, colon cancers, Alzheimer's and a host of other diseases (Gorman, Park, 2004). Human and mouse secreted interleukin-18 binding proteins (IL-18BPs) are known to block IL-18 activity, yet have no sequence similarity to membrane IL-18 receptors (Aizawa *et al*., 1999). Since overall aim of our work is to deduce minimal structures available for controlling the local and/or circulating IL-18 levels, we collected the amino acid sequences of various IL-18BPs from the NCBI database, aligned their IL-18 binding domains (IL-18BDs) and constructed a phylogenetic tree. Human, mouse and rat IL-18BP genes are split genes consisting of 5 to 7 exons (see GeneIDs: 10068, 16068 and 84388), whereas viral genes are intronless, indicating that they are retroposed from their host genomes (Xiang, Moss, 2003). The constructed phylogenetic tree suggested that at least two independent events created two different ancestral viral IL-18BP genes from the vertebrate lineage.

Methods

Using the amino acid sequence of human IL-18BP, we examined the homologous sequences by BLAST search at the NCBI and selected 18 sequences of IL-18BPs. Seven out of the 18 are coded by vertebrates, and the remaining 11 are coded by chordopoxviruses (Smith *et al*., 2000; Kim *et al*., 2000). Multiple alignment was performed with the CLUSTAL W (Thompson *et al*., 1994) and adjusted manually. Phylogenetic and molecular evolutionary analyses were conducted by the neighbor-joining (NJ) method (Saitou, Nei, 1989) in the MEGA version 2.2 (Kumar *et al*., 2001) .

Results and Discussion

We aligned amino acid sequences of the eighteen IL-18BDs (Fig. 1). Xiang Y. and Moss B. (2001a, b) determined the following 7 amino acid residues critical for high-affinity binding to human IL-18 within the human IL-18BD, *i.e*., 93-F, 97-Y, 99-L, 104-F, 106-E, 108-L and 114-E. Interestingly, our alignment revealed that all these 7 residues are widely conserved among the IL-18BDs examined.

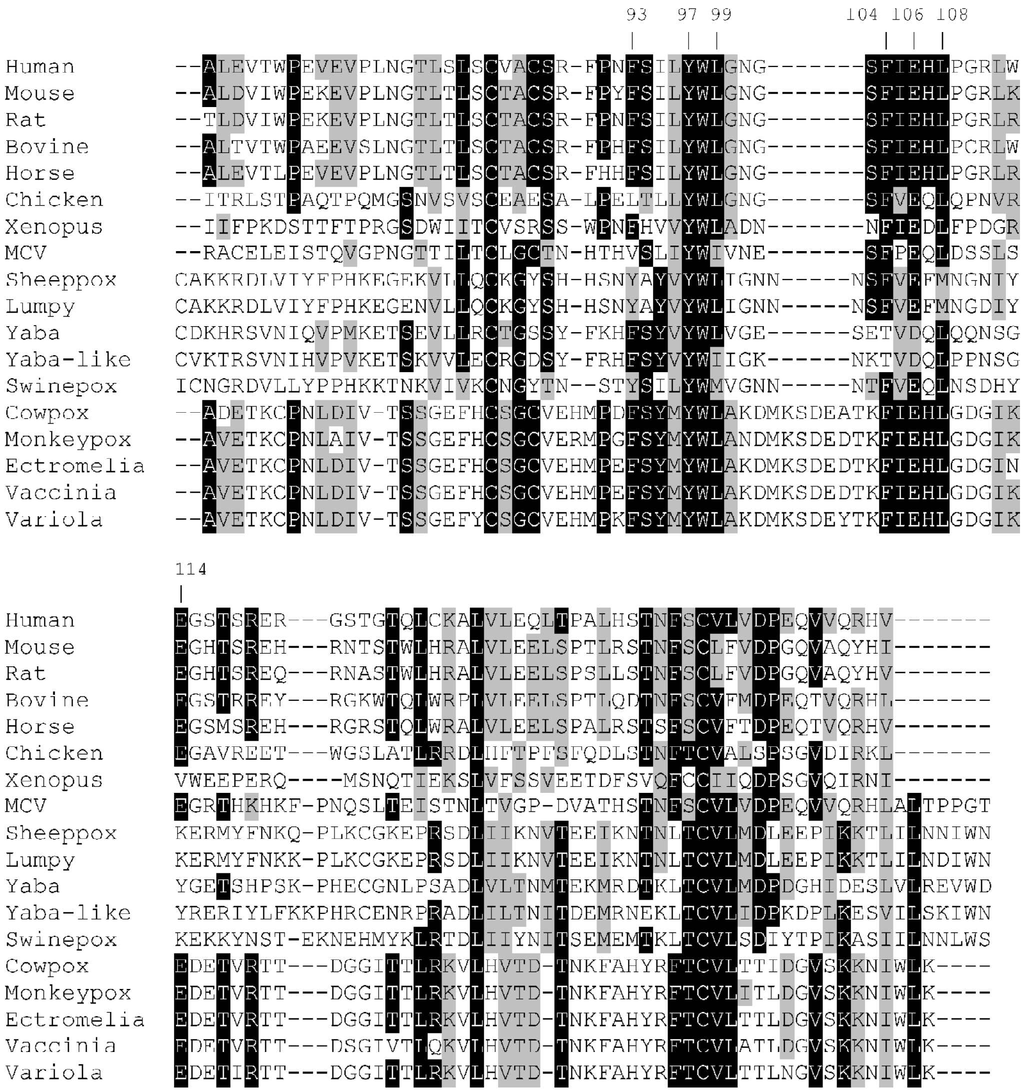

93 97 99 104 106 108
Human --ALEVTWPEVEVPLNGTLSLSCVACSR-FPNFSILYWLGNG-------SFIEHLPGRLW
Mouse --ALDVIWPEKEVPLNGTLTLSCTACSR-FPYFSILYWLGNG-------SFIEHLPGRLK
Rat --TLDVIWPEKEVPLNGTLTLSCTACSR-FPNFSILYWLGNG-------SFIEHLPGRLR
Bovine --ALTVTWPAEEVSLNGTLTLSCTACSR-FPHFSILYWLGNG-------SFIEHLPCRLW
Horse --ALEVTLPEVEVPLNGTLTLSCTACSR-FHHFSILYWLGNG-------SFIEHLPGRLR
Chicken --ITRLSTPAQTPQMGSNVSVSCEAESA-LPELTLLYWLGNG-------SFVEQLQPNVR
Xenopus --IIFPKDSTTFTPRGSDWIITCVSRSS-WPNFHVVYWLADN-------NFIEDLFPDGR
MCV --RACELEISTQVGPNGTTILTCLGCTN-HTHVSLIYWIVNE-------SFPEQLDSSLS
Sheeppox CAKKRDLVIYFPHKEGEKVLLQCKGYSH-HSNYAYVYWLIGNN-----NSFVEFMNGNIY
Lumpy CAKKRDLVIYFPHKEGENVLLQCKGYSH-HSNYAYVYWLIGNN-----NSFVEFMNGDIY
Yaba CDKHRSVNIQVPMKETSEVLLRCTGSSY-FKHFSYVYWLVGE------SETVDQLQQNSG
Yaba-like CVKTRSVNIHVPVKETSKVVLECRGDSY-FRHFSYVYWIIGK------NKTVDQLPPNSG
Swinepox ICNGRDVLLYPPHKKTNKVIVKCNGYTN--STYSILYWMVGNN-----NTFVEQLNSDHY
Cowpox --ADETKCPNLDIV-TSSGEFHCSGCVEHMPDFSYMYWLAKDMKSDEATKFIEHLGDGIK
Monkeypox --AVETKCPNLAIV-TSSGEFHCSGCVERMPGFSYMYWLANDMKSDEDTKFIEHLGDGIK
Ectromelia --AVETKCPNLDIV-TSSGEFHCSGCVEHMPEFSYMYWLAKDMKSDEDTKFIEHLGDGIN
Vaccinia --AVETKCPNLDIV-TSSGEFHCSGCVEHMPEFSYMYWLAKDMKSDEDTKFIEHLGDGIK
Variola --AVETKCPNLDIV-TSSGEFYCSGCVEHMPKFSYMYWLAKDMKSDEYTKFIEHLGDGIK
114
Human EGSTSRER---GSTGTQLCKALVLEQLTPALHSTNFSCVLVDPEQVVQRHV-------
Mouse EGHTSREH---RNTSTWLHRALVLEELSPTLRSTNFSCLFVDPGQVAQYHI-------
Rat EGHTSREQ---RNASTWLHRALVLEELSPSLLSTNFSCLFVDPGQVAQYHV-------
Bovine EGSTRREY---RGKWTQLWRPLVLEELSPTLQDTNFSCVFMDPEQTVQRHL-------
Horse EGSMSREH---RGRSTQLWRALVLEELSPALRSTSFSCVFTDPEQTVQRHV-------
Chicken EGAVREET---WGSLATLRRDLIIFTPFSFQDLSTNFTCVALSPSGVDIRKL-------
Xenopus VWEEPERQ----MSNQTIEKSLVFSSVEETDFSVQFCCIIQDPSGVQIRNI-------
MCV EGRTHKHKF-PNQSLTEISTNLTVGP-DVATHSTNFSCVLVDPEQVVQRHLALTPPGT
Sheeppox KERMYFNKQ-PLKCGKEPRSDLIIKNVTEEIKNTNLTCVLMDLEEPIKKTLILNNIWN
Lumpy KERMYFNKK-PLKCGKEPRSDLIIKNVTEEIKNTNLTCVLMDLEEPIKKTLILNDIWN
Yaba YGETSHPSK-PHECGNLPSADLVLTNMTEKMRDTKLTCVLMDPDGHIDESLVLREVWD
Yaba-like YRERIYLFKKPHRCENRPRADLILTNITDEMRNEKLTCVLIDPKDPLKESVILSKIWN
Swinepox KEKKYNST-EKNEHMYKLRTDLIIYNITSEMEMTKLTCVLSDIYTPIKASIILNNLWS
Cowpox EDETVRTT---DGGITTLRKVLHVTD-TNKFAHYRFTCVLTTIDGVSKKNIWLK----
Monkeypox EDETVRTT---DGGITTLRKVLHVTD-TNKFAHYRFTCVLITLDGVSKKNIWLK----
Ectromelia EDETVRTT---DGGITTLRKVLHVTD-TNKFAHYRFTCVLTTLDGVSKKNIWLK----
Vaccinia EDETVRTT---DSGIVTLQKVLHVTD-TNKFAHYRFTCVLATLDGVSKKNIWLK----
Variola EDETIRTT---DGGITTLRKVLHVTD-TNKFAHYRFTCVLTTLNGVSKKNIWLK----

Fig. 1. Multiple sequence alignment of IL-18 binding domains (IL-18BDs). Multiple alignment was performed as described in the **Methods**. A part of the human IL-18 binding protein (IL-18BP), from amino acid residues 64 to 161, together with the corresponding regions of the other IL-18BDs is aligned. The numbers above the alignment indicate the residue numbers of human IL-18BP critical for high-affinity binding to human IL-18; 93-F, 97-Y, 99-L, 104-F, 106-E, 108-L and 114-E. The dark and light shading indicate the residues that are more than 50 % identical and/or similar among the members, respectively. The accession numbers of sequences used are as follows: human (*Homo sapiens*, gi|10835224, isoform C), mouse (*Mus musculus*, gi|6754314, isoform C), rat (*Rattus norvegicus*, gi|16758106), bovine (*Bos taurus*, gi|29228534*), horse (*Equus caballus*, gi|31392907*), chicken (*Gallus gallus*, gi|32278657*), Xenopus (*Xenopus laevis*, gi|28245970*), MCV (Molluscum contagiosum virus, gi|9628986), Sheeppox (Sheeppox virus, gi|21492469), Lumpy (Lumpy skin disease virus, gi|15150454), Yaba (Yaba monkey tumor virus, gi|38229179), Yaba-like (Yaba-like disease virus, gi|12084997), Swinepox (Swinepox virus, gi|18640097), Cowpox (Cowpox virus, gi|20178392), Monkeypox (Monkeypox virus,gi|17974922), Ectromelia (Ectromelia virus, gi|7688160), Vaccinia (Vaccinia virus, gi|7688443) and Variola (Variola minor virus, gi|7514321). * indicates that the sequence was picked from a nucleotide database.

A constructed phylogenetic tree shown in Fig. 2 indicates that the chordopoxviruses can be separated into the three groups with high bootstrap supports: (i) MCV group, (ii) Vaccinia group and (iii) Yaba group, respectively. Apparently, the genetic distances within the Vaccinia group are smaller than those within the Yaba group.

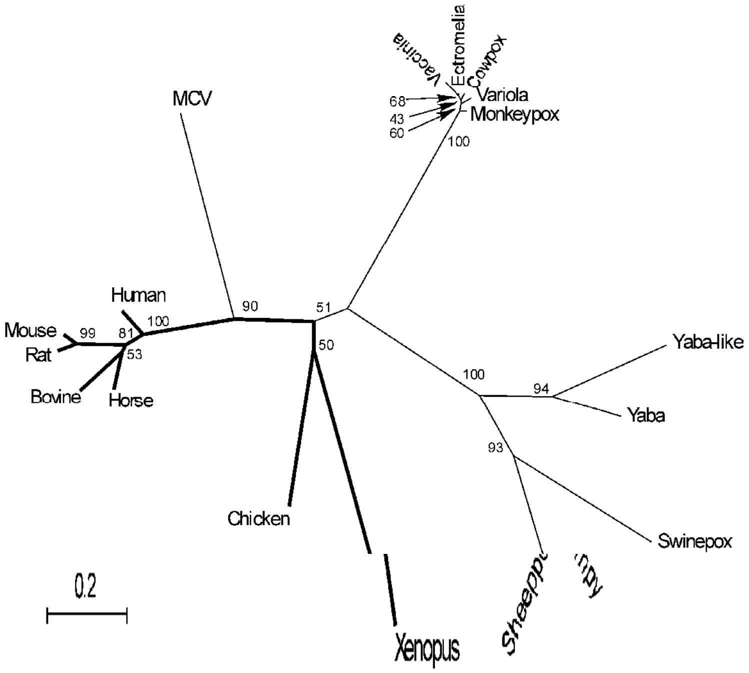


Fig. 2. A phylogenetic tree of IL-18 binding domains (IL-18BDs). Phylogeny was estimated using NJ method in the MEGA version 2.2. The bootstrap value obtained from 1000 replicates is shown at each branching point. Scale bar represents an estimate of the number of amino acid substitution per site. The solid line represents vertebrate lineage and the thin line, virus lineage, respectively. The abbreviations are as described in the legend to Fig. 1.

We believe that an ancestral viral IL-18BP gene should have been generated from a vertebrate genome by a retroposition event. The phylogenetic tree suggests that at least two independent events created the two different ancestral viral IL-18BP genes. The first event transferred a vertebrate IL-18BP gene to the genome of a common ancestral virus for the Vaccinia and Yaba groups, and the second event transferred the vertebrate IL-18BP gene to the ancestral genome of MCV. Recently, we found that proteins encoded by salamander (gi|45835077), trout (gi|24588628) and catfish (gi|20126236) show significant similarity to the human IL-18BP. However, we found that all these proteins show significantly higher similarity to interleukin-1 receptor type II (IL-1R-2) (data not shown), suggesting that IL-18BP and IL-1R-2, both have been evolved from a common ancestor. Study of the phylogenetic relationships of IL-18BP and IL-1R-2 will be our next project.

Acknowledgements

We are grateful to Dr. Etsuro Yamanishi, president of Hirakata Ryoikuen, for his supports and encouragements. We also wish to thank the National Center for Biotechnology Informatics, USA, for providing access to the network BLAST server, and the Center for Information Biology and DNA Data Bank of Japan, Shizuoka, for access to the CLUSTAL W analysis.

References

Aizawa Y. *et al*. Cloning and expression of interleukin-18 binding protein // FEBS Lett. 1999. V. 445. P. 338–342.

Gorman C., Park A. The fires within // Time. 2004. V. 4. P. 32–40.

Kim S.H. *et al*. Structural requirements of six naturally occurring isoforms of the IL-18 binding protein to inhibit IL-18 // Proc. Natl Acad. Sci. USA. 2000. V. 97. P. 1190–1195.

Kumar S. *et al*. MEGA2: molecular evolutionary genetics analysis software // Bioinformatics. 2001. V. 17. P. 1244–1245.

Saitou N., Nei M. The neighbor-joint method: a new method of reconstructing phylogenetic tree // Mol. Biol. Evol. 1987. V. 4. P. 406–425.

Smith V.P. *et al*. Ectromelia, vaccinia and cowpox viruses encode secreted interleukin-18-binding proteins // J. Gen. Virol. 2000. V. 81(Pt 5). P. 1223–1230.

Thompson J.D. *et al*. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice // Nucleic Acids Res. 1994. V. 22. P. 4673–4680.

Xiang Y., Moss B. Determination of the functional epitopes of human interleukin-18-binding protein by site-directed mutagenesis // J. Biol. Chem. 2001a. V. 276. P. 17380–17386.

Xiang Y., Moss B. Correspondence of the functional epitopes of poxvirus and human interleukin-18-binding proteins // J. Virol. 2001b. V. 75. P. 9947–9954.

Xiang Y., Moss B. Molluscum contagiosum virus interleukin-18 (IL-18) binding protein is secreted as a full-length form that binds cell surface glycosamino-glycans through the C-terminal tail and a furin-cleaved form with only the IL-18 binding domain // J. Virol. 2003. V. 77. P. 2623–2630.

SCANNING THE HUMAN GENOME FOR REGULATORY ISLANDS WITH PHYLOGENETIC FOOTPRINTING ALGORITHM

Wyrwicz L.S., Rychlewski L.*

Bioinformatics Laboratory, BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznan, Poland
* Corresponding author: e-mail: lucjan@bioinfo.pl

Keywords: *promoter, human genome, gene regulation, comparative genomics*

Summary

Motivation: The regulation of gene expression relies on various cis- and trans-acting elements. Even though the sequence of the human genome is known – the annotation of complete regulons is not possible, mostly due to the low information content of mean transcription factor binding site.

To locate regions likely to be involved in gene regulation we scanned the genomic sequences with phylogenetic footprinting propensities created from a set of aligned human and mouse promoters.

Results: The application of this algorithm, in addition to the intragenic sequence evolutionary conservation can help to create a reliable method aimed at annotation of proximal and distal promoter regions.

Availability: http://prosnp.bioinfo.pl/protest

Introduction

The regulation of gene expression relies on various cis- and trans-acting elements. Even though the sequence of human genome is known – the annotation of complete regulon is not possible, mostly due to the low information content of transcription factor binding site.

In the present study, we demonstrate the application of the phylogenetic footprinting algorithm in the identification of regulatory regions. We analyzed the mouse genome to model human promoter elements. Several authors insisted that the selection pressures applied during the separation from their last common ancestor allowed for the accumulation of sufficient mutations in regions with minimal effect on the regulation of expression, while constraining regions of regulatory importance (Pennacchio, 2001a, b).

In previous study (Wyrwicz, 2004) we have tested the conservation of short k-mers from five to nine residues in length. The comparison of 'sequence neighborhoods' (sets of short sequences that differ only by a single nucleotide) resulted in the development of a novel resource, which aims to characterize human promoters and asses the impact of polymorphisms on gene expression. Here we report the application of the previously developed regulation propensities for oligonucleotides in scanning for proximal (minimal) and distal promoter regions in human genome.

Methods

Genomic sequences and gene annotations. Human (build "hg16", June 2003) and mouse (build "mm3", Oct 2003) whole genome alignments were obtained from the Genome Browser (http://genome.ucsc.edu) site (Schwartz, 2003). The positions of 19,174 human genes from the Reference Sequence (Pruitt, 2001) project were retrieved from the Genome Browser data set. The splicing variants were excluded and 16,749 distinct genomic loci were analyzed in further steps.

Preparation of promoter alignments. Promoter sequences from human and mouse genes were extracted from the human-mouse genome alignments according to the location of their transcription starting site (TSS) relative to the respective human genes. We set the size of analyzed promoter sequence to 1000 base pairs (bp) upstream and 100 bp downstream from the TSS.

Sequence neighborhood. We scanned the promoter alignments for occurrences and evolutionary preservation (conservation) ratios of all motifs ranging from 5 to 9 residues. We treated k-mer as conserved when it occurred in both genomes at corresponding positions in promoters in not changed (mutated) form. We calculated conservation ratios of each sequence according to the equation:

$$c_i = C_i / T_i, \tag{1}$$

where c_i – is conservation ratio for motif i, C_i – conserved occurrences of motif i, T_i – total occurrence of motif i in the set of human promoters). For each motif a "sequence neighborhood" was constructed by including all motifs that differ by exactly one nucleotide. The number of sequences in the neighborhood set depended on the length of the motif ($N = 3 * n + 1$, where n is length of motif, N includes the starting motif). We used the Chi-square test to identify sequences with significantly different ratio of conservation (c_i) when compared to members of its neighborhood. The score for each sequence is described by Z-score calculated according to the equation:

$$Z = (c_i - c_{mean}) / SD_i, \tag{2}$$

where: c_{mean} – mean c for sequence neighborhood of i, SD_i – standard deviation of c for the neighborhood.

Genome annotation web-tool. The promoter propensities for k-mers were incorporated in the web-based tool – ProTest, available at URL: http://prosnp.bioinfo.pl/protest

The results are returned as cumulative promoter potentials for both DNA strands. The promoter potential score was calculated according to the equation:

$$P = \sum Z(k_{i,\, Z>j}) / f, \tag{3}$$

where: P – promoter potential in frame f; $Z(k_i)$ – Z score for k-mer 'i' (only k-mers with Z-score higher than j, where included), f – frame size. The tool allows modifying four parameters of scanning: a) frame size; b) k-mer library (5–9 bp); c) minimal Z-score for k-mer; d) minimal promoter potential score for consideration of region as regulatory.

The default settings are: 100bp frame, k-mers from 5–9 bp, Z-score > 2.2, P > 0.3.

The results can be visualized in several ways: graph, tab-delimited table and 'BED' genome annotation file format. The results stored in BED file format, allows us to compare the promoter predictions with human genome annotations prepared in the UCSC GenomeBrowser (http://genome.ucsc.edu/) (Kent, 2002).

The tool allows to annotate the genomic sequences up to 1 Mbp. Sequences can be provided by the user in FASTA format or preferably by submitting the location in the human genome (in format: chrA:B-C, where A stays for chromosome, B and C are the start and end positions respectively, positions refers to the build 16 of the human genome form July 2003).

The further analysis of identified putative regulatory regions can be performed with the ProSNP promoter analysis workbench (Wyrwicz, 2004). ProTest tool predictions are directly linked to the ProSNP workbench.

Results and Discussion

We have located putative regulatory regions in the human genome. The precomputed results for human chromosomes 20, 21 and 22 are available at URL: http://prosnp.bioinfo.pl/protest/chr2x.pl

The ProTest predictions performed for human chromosome 20, 21 and 22 revealed the concentrations of identified regulatory regions upstream from transcription start sites. The predictions coexist with potential regulatory regions identified with methods based on analysis of multiple genome alignments (human, mouse, rat; compare Fig.) (Gibbs, 2004; Schwart, 2003).

To validate the results we compared the positions of the identified regulatory regions with promoters annotated with FirstEF tool (Davuluri, 2001). Our method identified 38.3 % (1161 of 3027), 37.2 % (410 of 1101) and 42.7 % (1076 of 2518) of promoters annotated at human chromosomes 20, 21 and 22 respectively – mostly the regions located upstream from known genes (data not shown).

The further investigation of identified regions will focus on comparative predictions of human and mouse loci. This analysis will be performed for genomic sequences of genes with comparable expression profile.

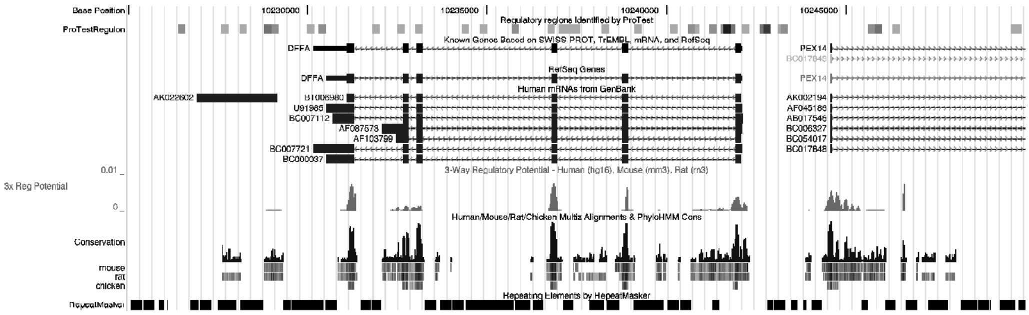


Fig. Results of ProSNP annotation of DFFA and PEX14 region in chromosome 1, visualized in UCSC Genome Browser. The regions identified by ProTest are marked as bars in upper row of graph. The intensivity of color correlates with score of ProSNP promoter potential. The ‘Conservation’ graph is created from BLASTZ multiple genome alignments (Schwartz, 2003). The ‘Reg potential’[Gibbs 2004] is the 3-way regulatory potential (RP) score, computed from alignments of human (hg16, Jul.’03), mouse (mm3, Feb. ’03) and rat (rn3, Jun.’03). RP scores compare frequencies of short alignment patterns between regulatory elements and neutral DNA.

Notice that the concentration of regions identified by ProSNP corresponds with 5’ends of genes, exons and conserved regulatory regions.

Acknowledgements

This work was supported by the EC 6FP and the MNiI within “BioSapiens” Network of Excellence (LHSG-CT-2003-503265).

References

Davuluri R.V., Grosse I., Zhang M.Q. Computational identification of promoters and first exons in the human genome // Nat Genet. 2001. V. 29(4). P. 412–7.

Gibbs R.A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution // Nature. 2004. V. 428(6982). P. 493–521.

Kent W.J., Sugnet C.W. *et al.* The Human Genome Browser at UCSC // Genome Res. 2002. V. 12. P. 996–1006.

Pennacchio L.A., Olivier M. *et al.* An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing // Science. 2001a. V. 294(5540). P. 169–173.

Pennacchio L.A., Rubin E.M. Genomic strategies to identify mammalian regulatory sequences // Nat. Rev. Genet. 2001b. V. 2(2). P. 100–109

Pruitt K.D., Maglott D.R. RefSeq and LocusLink: NCBI gene-centered resources // Nucleic Acids Res. 2001. V. 29(1). P. 137–140

Schwartz S., Kent W.J. *et al.* Human-mouse alignments with BLASTZ // Genome Res. 2003. V. 13(1). P. 103–107.

Wyrwicz L.S., Rychlewski L., Ostrowski J. ProSNP: phylogenetic footprinting in characterization of functional elements in human promoters // Nucl Acid Res. 2004, (submitted).

DATING POPULATION EXPANSION BASED ON STR VARIATION WITHIN Y-CHROMOSOME SNP-HAPLOGROUPS

Zhivotovsky L.A.

N.I. Vavilov Institute of General Genetics RAS, Moscow, Russia; e-mail: levzh@hotmail.com

Keywords: *human evolution, DNA markers, STRs, SNPs, dating, mathematical models, statistical methods*

Summary

Motivation: Dating divergence within- and between DNA-lineages is of great importance for revealing historic information from observed DNA variation in human populations. The increasing amount of data on Y-chromosome SNPs and STRs require adequate methods for estimation of population expansion time. Such methods should necessarily involve mathematical models that describe evolution of DNA variation, and statistical methods to estimate the dates of population events.

Results: Dynamic models of population divergence under mutation, genetic drift and migration are analyzed. Estimates of the dynamic parameters are introduced and applied to several data sets on worldwide human populations.

Introduction

Increasing attention has recently been paid to microsatellite variation within Y-chromosome haplogroups defined by binary polymorphisms, such as SNPs (single nucleotide polymorphisms), or, as a general term, UEPs – unique event polymorphisms (Underhill *et al.*, 1996; Zerjal *et al.*, 1997; Kayser *et al.*, 2000a; de Knijff, 2000), many of which are specific to populations related through their recent or past history (Underhill *et al.*, 2000; Hammer *et al.*, 2001; Y Chromosome Consortium, 2002). Microsatellites, or short tandem repeat (STR) polymorphisms, are abundant in the human genome, can be easily genotyped and scored, and thus have become a useful tool for the elucidation of human population history and for forensic purposes. The mutation rate at Y-chromosome STRs is important for calibration of the molecular clock in evolutionary studies and has been estimated (Heyer *et al.*, 1997; Forster *et al.*, 2000; Zhivotovsky *et al.*, 2004). Similar linked SNP and STR haplotypes are also available in autosomes (Mountain *et al.*, 2002). How this information can be used for dating historic population events is illustrated in this paper.

Dynamic models

Variation at STR loci is amenable to the use of population statistics that treat allele repeat scores (the number of repeats) as a quantitative trait (Goldstein *et al.*, 1995a, b; Slatkin, 1995; Zhivotovsky, Feldman, 1995). Among them is a specific measure of genetic distance, the squared difference in the number of repeats, which is linear with time since divergence (Goldstein *et al.*, 1995b; Zhivotovsky, Feldman, 1995). However, this property is fulfilled only if populations under divergence are at genetic equilibrium, have constant equal sizes and are not subject to gene flows; therefore a different, though related genetic distance has been suggested (Zhivotovsky, 2001) that is relatively robust to these demographic processes. We model these processes and a modified distance to apply it for Y chromosome STRs within lineages defined by SNPs.

Statistical method of dating

The age of STR variation within each haplogroup can estimated as the average squared difference (ASD_0) in the number of repeats between all current chromosomes of a sample and the founder (assumed to be modal) haplotype divided by $w = 6.9 \times 10$-4 per 25 years (Zhivotovsky *et al.*, 2004). The upper bound for expansion time (the time of divergence of populations) is suggested to be calculated using T_D (Zhivotovsky, 2001) and assuming an STR-variance in repeat scores at the beginning of population separation (V_0) equal to zero. The lower bound is calculated as T_D, with V_0 taken as a predicted value of the within-population STR-variance prior to population split; the latter was computed as a linear approximation of the within-population variance in repeat scores as a function of time. We investigate how these estimates evolve using the model just described above, and evaluate how informative they are under uncertainties in demographic parameters.

Population data. Results and Discussion

For illustration of our approach we use data on populations with various histories. Among those are Polynesian populations (Maoris, Cook Islanders, and Samoans) whose Y-chromosome lineages reflect Polynesian origin in Melanesia and eastern/southeastern Asia, in particular lineage C2 characterized by mutations at RPS4Y711 and M38 marked additionally with the mutation M208 (Su *et al.*, 2000; Kayser *et al.*, 2000a; Underhill *et al.*, 2001). The Gypsy populations from Bulgaria were analyzed with a Y chromosome lineage defined by mutation M82, which is derived from the Indian subcontinent and is exceedingly rare in Europe (Semino *et al.*, 2000; Underhill *et al.*, 2000; Gresham *et al.*, 2001). The Bantu expansion was investigated using the E3a7-M191 haplogroup, which occurs at high frequency in the Bantu populations, with traces in other, non-Bantu-speaking groups from sub-Saharan Africa (Cruciani *et al.*, 2002). Haplogroups E and J was investigated in samples from Europe and the Mediterranean and also from Africa and Asia, which are distributed differentially within the Near East, North Africa and Europe probably associating to the diffusion of Arab people or reflecting the spread of Anatolian farmers, or tracing the subsequent diffusion of people from the southern Balkans to the West (Semino *et al.*, 2004). Around 20 Native American, 28 Asian, and 5 European populations (including 342 Amerind speakers, 186 Na-Dene speakers, and 60 Aleut-Eskimo speakers) were used to investigate the origins of Native American paternal lineages based on SNP analysis of three major haplogroups C, Q, and R, accounted for almost all Native American Y chromosomes (Zegura *et al.*, 2004).

The data illustrate how the statistical procedure works and how the estimates correspond to available historic and archaeological records.

Acknowledgements

This work was partially supported by RAS programmes “Genetic aspects of the Biosphere evolution” and “Genofond dynamics” and grants RFBR 04-04-48639 and NIH R03 TW005540.

References

Cruciani F., Santolamazza P., Shen P., Macaulay V., Moral P., Olckers A., Modiano D., Holmes S., Destro-Bisol G., Coia V., Wallace D.C., Oefner P.J., Torroni A., Cavalli-Sforza L.L., Scozzari R., Underhill P.A. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes // Amer. J. Hum. Genet. 2002. V. 70. P. 1197–1214.

De Knijff P. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome // Amer. J. Hum. Genet. 2000. V. 67. P. 1055–1061.

Forster P., Röhl A., Lünnemann P., Brinkmann C., Zerijal T., Tyler-Smith Ch., Brinkmann B. A short tandem repeat-based phylogeny for the human Y chromosome // Amer. J. Hum. Genet. 2000. V. 67. P. 182–196.

Goldstein D.B., Linares A.R., Cavalli-Sforza L.L., Feldman M.W. An evaluation of genetic distances for use with microsatellite loci // Genetics. 1995a. V. 139. P. 463–471.

Goldstein D.B., Linares A.R., Cavalli-Sforza L.L., Feldman M.W. Genetic absolute dating based on microsatellites and the origin of modern humans // Proc. Natl Acad. Sci. USA. 1995b. V. 92. P. 6723–6727.

Gresham D., Morar B., Underhill P.A., Passarino G., Lin A.A., Wise Ch., Angelicheva D., Calafell F., Oefner P.J., Shen P., Tournev I., de Pablo R., Kuinskas V., Perez-Lezaun A., Marushiakova E., Popov V., Kalaydjieva L. Origins and divergence of the Roma (Gypsies) // Amer. J. Hum. Genet. 2001. V. 69. P. 1314–1331.

Heyer E., Puymirat J., Dietjes P., Bakker E., de Knijff P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees // Hum. Mol. Genet. 1997. V. 6. P. 799–803.

Kayser M., Brauer S., Weiss G., Schiefenhövel W., Underhill P., Shen P., Oefner P., Tommaseo-Ponzetta M., Stoneking M. Reduced Y-chromosome, but not mtDNA, diversity in human populations from West New Guinea // Amer. J. Hum. Genet. 2003. V. 72. P. 281–302.

Kayser M., Brauer S., Weiss G., Underhill P.A., Roewer L., Schiefenhovel W., Stoneking M. Melanesian origin of Polynesian Y chromosomes // Current Biol. 2000a. V. 10. P. 1237–1246.

Kayser M., Roewer L., Hedman M., Henke L., Henke J., Brauer S., Krüger K., Krawczak M., Nagy M., Dobosz T., Szibor R., de Knijff P., Sajantila A. Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs // Amer. J. Hum. Genet. 2000b. V. 66. P. 1580–1588.

Mountain J.L., Knight A., Jobin M., Gignoux C., Miller A., Lin A.A., Underhill P.A. SNPSTRs: Empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes // Genome Res. 2002. V. 12. P. 1766–1722.

Semino O., Magri Ch., Benuzzi G., Lin A.A., Al-Zahery N., Battaglia V., Maccioni L.a, Triantaphyllidis C., Shen P., Oefner P.J., Zhivotovsky L.A., King R., Torroni A., Cavalli-Sforza L.L., Underhill P.A., Santachiara-Benerecetti A.S. Origin, diffusion and differentiation of Y-chromosome haplogroups E and J: inferences on the Neolithization of Europe and later migratory events in the Mediterranean area // Amer. J. Hum. Genet. 2004. (In press).

Slatkin M. A measure of population subdivision based on microsatellite allele frequencies // Genetics. 1995. V. 139. P. 457–462.

Su B., Jin L., Underhill P.A., Martinson J., Saha N., McGarvey S.T., Shriver M.D., Chu J., Oefner P., Chakraborty R., Deka R. Polynesian origins: insights from the Y chromosome // Proc. Natl Acad. Sci. USA. 2000. V. 97. P. 8225–8228.

Underhill P.A., Passarino G., Lin A.A., Marzuki S., Oefner P.J., Cavalli-Sforza L.L., Chambers G.K. Maori origins, Y-chromosome haplotypes and implications for human history in the Pacific // Human Mutation. 2001. V. 17. P. 271–280.

Underhill P.A., Shen P., Lin A.A., Jin L., Passarino G., Yang W.H., Kauffman E., Bonne-Tamir B., Bertranpetit J., Francalacci P., Ibrahim M., Jenkins T., Kidd J.R., Mehdi S.Q., Seielstad M.T., Wells R.S., Piazza A., Davis R.W., Feldman M.W., Cavalli-Sforza L.L., Oefner P.J. Y chromosome sequence variation and the history of human populations // Nat. Genet. 2000. V. 26. P. 358–361.

Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups // Genome Res. 2002. V. 12. P. 339–348.

Zerjal T., Dashnyam B., Pandya A., Kayser M., Roewer L., Santos F., Schiefenhövel W., Fretwell N., Jobling M.A., Harihara S., Shimizu K., Semjidmaa D., Sajantila A., Salo P., Crawford M.H., Evgrafov O., Tyler-Smith C. Genetic relationships of Asians and northern Europeans revealed by Y-chromosomal DNA analysis // Amer. J. Hum. Genet. 1997. V. 60. P. 1174–1183.

Zegura S.L., Karafet T.M., Zhivotovsky L.A., Hammer M.F. High resolution SNPs and microsatellite haplotypes point to a single, recent entry of native American Y chromosomes into the Americas // Mol. Biol. & Evol. 2004. V. 21. P. 164–175.

Zhivotovsky L.A. Estimating divergence time with the use of microsatellite genetic distances: impacts of population growth and gene flow // Mol. Biol. Evol. 2001. V. 18. P. 700–709.

Zhivotovsky L.A., Feldman M.W. Microsatellite variability and genetic distances // Proc. Natl Acad. Sci. USA. 1995. V. 17. P. 11549–11552.

Zhivotovsky L.A., Underhill P.A., Cinnioglu C., Kayser M., Morar B., Kivisild T., Scozzari R., Cruciani F., Destro-Bisol G., Spedini G., Chambers G.K., Herrera R.J., Yong K.K., Gresham D., Tournev I., Feldman M.W., Kalaydjieva L. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time // Amer. J. Human Genet. 2004. V. 74. P. 50–61.

AN EVOLUTIONARY LINEAGE FOR INTRON LOSS/GAIN IN FIVE EUKARYOTIC GENOMES

Zhou Z., Kwoh C.K.*

Bioinformatics Research Centre, School of CE, Nanyang Technological University, Singapore 639798
* Corresponding author: e-mail: ezlzhou@ntu.edu.sg

Keywords: *eukaryotic genomes, sequence conservation, gene structure, intron gain, intron loss*

Summary

Comparative genome analysis indicates that there is a sizeable amount of protein sequence conservation between any two known eukaryotic genomes. However, the gene structures of orthologous proteins may or may not be conserved. We examine five model organisms (human, fruit fly, weed, worm, and yeast) for negative correspondence between conservation in gene structure and protein sequence. Here, we report for the first time a minimum set of 64 orthologous proteins that exhibit significant sequence conservation between the five model organisms, whilst the corresponding gene structures are found to be arbitrarily divergent. We find that the majority of human genes represented within this orthologous group have gained introns, followed in order by weed, worm, fly and yeast. This minimal dataset provides an opportunity to describe the random phenomenon of intron gain/loss during eukaryotic evolution.

Introduction

The complete genome sequence data is available for six eukaryotic species (Adams *et al.*, 2000; Goffeau *et al.*, 1997; The *C. elegans* Sequencing Consortium, 1998) and the data provides the possibility of identifying sequence conservation among them by comparative genome analysis. Knowledge on the extent to which gene structure is conserved through vertical gene transfer will guide reconstruction and inference of evolutionary history, and has direct bearing on any idea about the mechanisms of diversification and speciation. Efforts are underway to understand horizontal gene flow among known microbial genome species. Horizontal gene transfer is generally studied by estimating sequence homology and gene characteristics of orthologous genes in different species. Fifty-one cases of gene fusions represented in at least two of the three primary kingdoms (Bacteria, Archaea and Eukaryota) are recently reported. The euGenes genome information system provides data to study genome relationships across several eukaryotic species (Gilbert, 2002). This system reports nearly 6-44 % gene conservation between any two species among human, weed, worm, fly and yeast (Gilbert, 2002). For example, 44 % of genes in the fruit fly genome may have some homology to human genes while 19 % of genes in the human genome may have some homology to fly genes and so on. Although, there is a reasonable amount of gene product conservation between any two species in eukaryotes, their exon-intron structures would be expected to randomly diverge due to intron loss/gain within evolutionary time. The observed consensus thus far argues for a preponderance of intron loss. However, recent cases of spliceosomal intron gain are also reported. We believe that isolated cases of either intron loss or gain do not provide a sufficient basis by which the evolutionary lineage of eukaryotic intron loss or gain can be reconstructed. The availability of complete genome sequences for human, weed, worm, fruit fly and yeast makes it possible to study the phenomenon of intron loss or gain in the ancestry of these eukaryotic species by the systematic analysis of their individual exon-intron structures. Here, we report an evolutionary lineage for intron evolution among these five eukaryotic model organisms.

Method

The protein sequences encoded by intronless (Sakharkar *et al*., 2002) and intron-containing genes (Sakharkar *et al*., 2000) for human, weed, worm, fly and yeast were used to identify the list of sequences that are conserved across all the 5 species. The five organisms that are compared all together in this study are *Arabidopsis thaliana* (AT), *Caenorhabditis elegans* (CE), *Drosophila melanogaster* (DM), *Homo sapiens* (HS), and *Saccharomyces cerevisiae* (SC). We identified 2,773 intronless (IL) and 9,701 intron-containing (IC) protein sequences for HS; 6,588 IL and 25,497 IC for AT; 1,447 IL and 20,140 IC for CE; 3,837 IL and 13,737 IC for DM; 5,115 IL and 468 for SC from the intronless (Sakharkar *et al*., 2002) and intron-containing dataset (Sakharkar *et al*., 2000). The redundant sequences in each of the 10 sequence sets were removed at 40 % sequence identity over at least 70 % of the sequences using the program CD-HIT (Li *et al*., 2001). The non-redundant sequence sets contain 785 IL and 4,568 IC for HS; 3,542 IL and 12,339 IC for AT; 706 IL and 14,506 IC for CE; 2,813 IL and 9,457 IC for DM and 3,676 IC and 180 IL for SC. All the non-redundant IL and IC protein sequences from the five species were clustered again at 40 % identity level over at least 70 % of the sequences using the CD-HIT (Li *et al*., 2001) and thereby we identified a representative set of 64 protein sequences conserved across all the 5 species. We assigned corresponding gene structure information to each sequence using the "CDS" entry in the GenBank FEATURES (Sakharkar *et al*., 2002; Sakharkar *et al*., 2000). For the clusters that contain obvious paralogs, higher preference was given to those sequences within the clusters with lowest sequence identity by the elimination of the paralogs which bore the greatest similarity to the cluster. The number of such paralogs eliminated by this process is very small and does not affect the conclusion drawn in this article.

Results and Discussion

Genome wide comparative analysis between pairs of genomes shows considerable protein sequence homology among them (The *C. elegans* Sequencing Consortium, 1998). However, we believe that there will be no significant exon-intron structure similarity among genes corresponding to orthologous protein sequences. This is mainly due to a phenomenon called intron gain or loss during evolution and this process is traditionally described as random. Knowledge on this highly unpredictable phenomenon is important for the understanding of eukaryotic gene evolution. Here, we construct an evolutionary lineage for intron loss and gain among human, weed, worm, fly and yeast using a set of protein sequences identified as conserved in all of them.

We used GenBank genomic data (release 128) to identify protein sequences that are conserved across HS, AT, CE, DM and SC (see Methods for details). The procedure results in a set of protein sequences that are conserved among these genome species in a non-redundant manner. The representative sequences used in this analysis for each of these genomes have less than or equal to 40 % sequence identity among themselves. Thus, we identified 64 clusters whose protein sequences are conserved across these organisms. When their corresponding exon-intron structures are mapped to each protein sequence in each genome species they are found to be randomly divergent between them. It should be noted that names and functions of many genes in these 64 clusters still remain undetermined. However, putative exon-intron structures are assigned to each of these genes by prediction technique. Previous reports suggest that two organisms share a sizeable amount of homology when compared two at a time (The *C. elegans* Sequencing Consortium, 1998). But the homology considerably drops when more than two genomes are compared all together. Therefore, a set of protein sequences homologous between two genomes need not be always homologous between any other two genomes. It is because of this reason we could identify only a limited number of 64 protein sequences being conserved among all the five eukaryotic species. This set constitutes only a small proportion of the genome. Universal cellular processes require many more than 64 genes implying that the same functions can be performed by non-homologous genes in different species. If we assume that each sequence in the dataset represents a unique fold in the genome then our result suggest that there are not many structural folds common among distant eukaryotic genome species. This implies the possibility of different evolutionary lineage for each of the five eukaryotic genome species.

The conserved protein sequences among genomes have randomly divergent gene structures between them. A single parameter, the number of introns (N) in these gene clusters varies from 0 to as many as 34 in different species. Intron loss and gain in genes from different eukaryotic genomes is highly random and this randomness may contribute towards the observed gene structure divergence in genes. Using exon-intron structural information in the 64 clusters we propose a lineage for intron gain/loss during eukaryotic genome evolution. Figure, is an illustration for intron loss/gain between SC, DM, CE, AT and HS. The number adjacent to the arrow indicates the percentage sequences that have gained or lost at least one intron between the species connected by the arrow. The solid line arrows indicate the direction of intron gain while the dot line arrows indicate the direction of intron loss between them. When a pathway is constructed connecting all the five species such that the arrow heads are always in the forward direction, the lineage starting in the order yeast-fly-worm-weed-human fits the criteria well. This lineage is indicated by a freeform dash line in Figure. The results suggest that intron gain might have possibly taken place from yeast to human, through fly, worm and weed. Similarly, intron loss would have taken place in the exact reverse order. It is clear that either intron loss or gain or both have occurred during eukaryotic evolution. What is unclear is that which among the three processes might have taken place most during intron evolution? In Figure, it is found that humans mostly gained introns compared to weed, worm, fly and yeast during evolution; weeds gained introns compared to worm, fly, and yeast, but lost introns compared to human; worms gained introns compared to fly and yeast, but lost introns compared to weed and human; fly gained introns compared to yeast, but lost introns compared to weed, worm and human; yeast lost introns compared to all the others. This strongly suggests that evolution either favored addition or deletion of introns. The

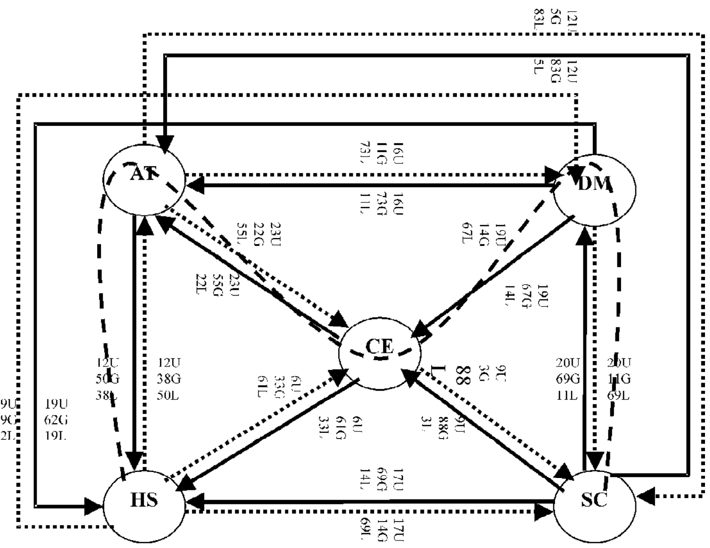


Fig. Percentages of conserved proteins resulting from multiple comparison among HS, AT, CE, DM and SC are shown. The organisms used in the analysis are shown in circles. Numbers that are adjacent to the arrows indicate the percentage of proteins that were found to show gain (G) or loss (L) of introns between the two organisms marked. U indicates the percentage of proteins that neither showed loss nor showed gain in the number of introns in their genes. The arrow-head shows the direction of intron evolution. Solid lines indicate the direction for intron gain and while, dot lines indicate the direction for intron loss. The freeform dash line shows the constructed lienage for both intron loss and intron gain with direction in the exactly reverse order. The freeform dash line connects all the five organisms such that the arrow-head connection are continuous and unindirectional.

difference between percentage intron loss and gain is least between HS and AT while it is most between SC and DM. This implies that humans and weeds are more closely related compared to others in the context of intron evolution. The results imply that intron loss and gain in the evolutionary tree is certainly not unidirectional despite the observation that some species largely tend to gain more introns and others mostly tend to lose more introns.

References

Adams M.D. *et al*. The genome sequence of *Drosophila melanogaster* // Science. 2000. V. 287. P. 2185–2195.

Gilbert D.G. EuGenes: A eukaryote genome information system // Nucleic Acids Res. 2002. V. 30. P. 145–148.

Goffeau A. *et al*. The yeast genome directory // Nature. 1997. V. 387. (suppl.) 5.

Li W., Jaroszewski L., Godzik A. Clustering of highly homologous sequences to reduce the size of large protein database // Bioinformatics. 2001. V. 17. P. 282–283.

Sakharkar M.K., Kangueane P., Petrov D.A., Kolaskar A.S., Subbiah S. SEGE: A database on "intronless/ single exonic" genes from eukaryotes // Bioinformatics. 2002. V. 18. P. 1266–1267.

Sakharkar M.K., Long M., Tan T.W., de Souza S.J. ExInt: an Exon Intron Database // Nucleic Acids Res. 2000. V. 28. P. 191–192.

The *C. elegan*s Sequencing Consortium. Genome sequence of the nematode *C. elegans*: a platform for investigating biology // Science. 1998. V. 282. P. 2012–2018.

BGRS 2004

NEW APPROACHES
TO ANALYSIS
OF BIOMOLECULAR DATA
AND PROCESSES

A FAST PROCEDURE FOR MODELING OF PROTEASOMAL PROTEIN DEGRADATION *IN VITRO*

Antyufeev V.S.[1], **Nikolaev S.V.**[2]*

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

Keywords: *proteasome, protein degradation, mathematical model, computer analysis*

Summary

Motivation: Modeling of complex processes with stochastic elements frequently involves computer-based simulation models. Calculation of the statistical characteristics (parameters) of the modeled processes often is time consuming. For this reason, we attempted to analytically calculate the distribution of the required parameters of the complex processes.

Results: A simple model for proteasomal proteolysis has been constructed as a Markov chain. Formulas were derived for estimating the distribution of the fragments resulting from proteolysis. To do it, we built the appropriate phase space which allowed us to transform the modeled process into Markov chain, and to calculate the limit probability of its absorbing states. This enabled us to reduce by hundreds of times the time for modeling of the proteolytic process and to achieve good calculation accuracy in an acceptable time.

Introduction

A living cell contains a variety of proteins with a broad range of lifetime. If a protein exists, but the cell does not need it any longer, the cell gets rid of it by proteolysis (degradation, or enzymatic cleavage of proteins). About twenty years or so ago, a high molecular protein complex working as a proteolytic enzyme was found in the cell. The proteasome is cylinder-shaped with a channel along its axis. The channel diameter suffices to accommodate the protein molecule within the channel and to allow its movement. The "cleaving unit" of the proteasome is located at the inner surface of the channel, and it can split the protein molecule or its fragments between the amino acids.

Many details of the protein degradation process remain unknown. There is a host of intriguing questions: Does protein degradation start from the N- or the C-terminus? Is degradation processive, or the proteasome-protein complex keeps dissociating-reassociating in a protein undergoing degradation? What mechanism provides the movement of a degrading protein relative to a proteasome? The relevant experiments are carried out *in vitro* using 20S proteasomes and unfolded proteins (a protein with a linear spatial structure or a protein unfolded by a denaturating agent). Apparently, the protein substrate-proteasome relative movement is not provided by a known molecular motor driven by the ATP cleavage. The distribution of the length of the resulting fragments is described by a unimodal curve with a maximum in the range of 6–10 amino acids (Kisselev *et.al.*, 1998).

Our goal was to model proteasomal proteolysis. Attempts were made to answer questions, such as: Can the diffusive relative protein substrate-proteasome movement provide the splitting of the entire protein molecule into fragments? What cutting parameters produce a distribution of fragment length similar to the one experimentally observed?

The developed simulation model took too much time for computer-based experiments aimed at finding the distribution the fragments resulting from the proteolysis. For this reason, we turned to derivation of formulas expressing the distribution of the characteristics of the process.

Model

The following model of the proteasomal protein degradation process is considered.

We assume that there are numerous protein copies and some proteasomes in a solution at the beginning of the degradation process. As protein degradation proceeds, the fragments of protein copies appear in the solution. To model the degradation process, the points between neighboring amino acids in a protein sequence are considered rather than the amino acids as such. Every point is a possible point of protein molecule cleavage.

Because there are no universally recognized models of protein substrate movement relative to the proteasome, we postulate that this movement is a one-dimentional diffusion effected by thermal Brownian movement of the protein substrate along the proteasomal channel. During such Brownian movement, intermolecular interaction causes random glueing of the amino acids of the protein substrate and the ones of the proteasome. This glueing is the necessary but not a sufficient condition for proteasome to cleave the substrate molecule into two fragments. We suggest that the proteasome may cleave a protein or a protein fragment only if there are substrate-proteasome glueings at both sides of the "cleaving unit" of the proteasome, and the cleavage is random.

Algorithm for modeling

First we recall the basic steps of the proposed model. Suppose there is a pool of N protein fragments in a solution at a time. A fragment of the pool inputs in a proteasome with the probability $1/(2N)$ (the coefficient 1/2 reflects the random choice of the N- or C-terminus of a fragment), and moves randomly in the proteasomal channel along it axis. During this movement, the fragment may be cleaved or it may exit from the proteasomal channel, being not cleaved. It is noteworthy that modeling of a fragment movement relative to the proteasome is equivalent to modeling of the movement of the proteasomal "cleaving unit" relative to the fragment. Modeling of the entire process is stepwise: 1. Choice of a number and of a terminus of the fragment, which is set to be in contact with the proteasome. 2. Model of random discrete steps of the proteasomal "cleaving unit" in either direction along the chosen fragment until the fragment becomes detached. 3. Model of glueing of two or more substrate amino acids to the proteasome (see the model description above). 4. If two or more substrate amino acids fixed to the proteasome at both sides of the proteasomal "cleaving unit", then the division of the fragment into two fragments at the point under the "cleaving unit" is modeled. 5. If the fragment is cleaved into two smaller fragments, then the fragment that would leave the proteasome and the one that would remain in it are modeled. The detached fragment joins the substrate pool. Run the algorithm from point 2 with the remaining fragment. 6. If a fragment will not be divided during its movement through the proteasomal channel and the fragment detaches from the proteasome, then the fragment joins the substrate pool. Return to point 1 of the algorithm.

The choice of the division point is two-step: (i) the choice of the fragment to be divided, and (ii) of the division point in the preferred fragment. The fragments in the pool and the points of the possible division in the fragments are numbered. Then, the task is to sequentially model two discrete random values: the divided fragment number, and, next, the division point number. Once the discrete distributions are found, the time for modeling of the process in its entirety becomes much reduced. The two random processes can be expressed as absorbing Markov chains (Kemeny, Snell, 1969). For this purpose, one has to build a phase space of the process involving absorbing and non-absorbing states, transition and absorption probabilities. Let us consider modeling of the fragments resulting from division. The phase space consists of fictitious elements of the $[k,a,s]$ type, where k relates to the fragment number, a to the absorbance index; a=1 or 0, depending on whether the k-th fragment will be divided or not respectively; s denotes the particular (the right or left) side of the fragment penetrating into the proteasomal channel (note, after a next fragment has been divided, the phase space changes, and the next division points are chosen in it; this means that the entire process is not Markov chain. Because each subsequent step depends on all the preceding, the choice of the next division point depends on how they all were chosen). Suppose that there are *N* fragments, the phase space consists of *4N* elements.

The transition matrix of the chain falls into blocks:

$$P = \begin{bmatrix} A & B \\ 0 & E \end{bmatrix}.$$

All the squarte submatrices are of size $2N \times 2N$, E – unitary matrix,

$$A = \begin{bmatrix} p_1 q_1 & p_2 q_1 & \cdots & p_{2N} q_1 \\ p_1 q_2 & p_2 q_2 & \cdots & p_{2N} q_2 \\ \vdots & \vdots & \vdots & \vdots \\ p_1 q_{2N} & p_2 q_{2N} & \cdots & p_{2N} q_{2N} \end{bmatrix} \quad B = \begin{bmatrix} 1-q_1 & & & \\ & 1-q_2 & & \\ & & \ddots & \\ & & & 1-q_{2N} \end{bmatrix}.$$

Here, $(1 - q_i)$ stands for the probability of division of *i*–th fragment, all of the $p_i = 1/(2N)$.

The idea is to calculate the limit distribution for the probabilities of the absorbing states for the chain. The vector of the limit distribution is $u^* = \lim_{n \to \infty} uP^n$

By induction, we prove that $P^n = \begin{pmatrix} A^n & (E + A + \ldots + A^{n-1})B \\ 0 & E \end{pmatrix}.$

Since $\|A\|_1 < 1$, $\lim_{n \to \infty} (E + A + \ldots + A^{n-1}) = (E - A)^{-1}$.

Hence $P^* = \begin{pmatrix} 0 & (E - A)^{-1} B \\ 0 & E \end{pmatrix}$, and $uP^* = u \begin{pmatrix} 0 & (E - A)^{-1} B \\ 0 & E \end{pmatrix}.$

To calculate the above limit distribution, the following steps should be performed: (i) calculation of the vector *x*, as a solution of the linear equation set $(E - A)\, x = u$, where *u* – vector of the initial probability distribution, and (ii) calculation of the vector *xB*. In addition, we found a formal solution of the linear equation set. Thus, calculation of the distribution and modeling of the discrete random value with known distribution can be used instead of computer simulation of the fragment choice. The calculation of the probability distribution for the division points in the fragment is similar. Thus, the time consuming algorithm 1–7 can be replaced by fast calculation, followed by modeling of the discrete distributions. The consequence of this improvement is a decrease in the calculation time by two orders of magnitude.

Results and Discussion

The appropriate phase space is the major result of the work. The expansion of the phase space of the process by including the "artificial" elements, such as [*k,a,s*], along with the "usual" elements (such as fragment and division point numbers) allowed us to express the random process as an absorbing Markov chain. In this way, the limit probabilities of the absorbing states were calculated. Based on the defined probabilities, the modeling time was reduced by two orders of magnitude, and calculation accuracy was achieved in an acceptable time.

Acknowledgements

The work is supported by Russian Ministry of Industry, Sciences and Technologies (grant No. 43.073.1.1.1501), RAS Presidium Program "Molecular and Cellular Biology" (No. 10.4).

References

Kemeny J., Snell J. Finite Markov chains. The University Series in Undergraduate Mathematics, Dartmouth College. 1969.

Kisselev A.F., Akopian T.N., Goldberg A.L. Range of sizes of peptide products generated during degradation of different proteins by archaeal proteasomes // J. Bio. Chem. 1998. V. 273(4). P. 1982–1989.

A SYSTEM FOR ON-LINE PROCESSING OF IMAGES OF GENE EXPRESSION PATTERNS

Blagov M.S.*, Poustelnikova E.G., Pisarev A.S., Myasnikova E.M., Samsonova M.G.

St.Petersburg State Polytechnical University, St.Petersburg, Russia
* Corresponding author: e-mail: blagov@spbcas.ru

Keywords: *on-line processing, gene expression patterns, relational database, confocal scanning microscopy, image processing*

Summary

Motivation: The development of software for quantification and analysis of gene expression *in situ* is an important task for bioinformatics.

Results: A software system to provide access and process images of gene expression patterns has been developed. The images can be stored in a relational database as BLOBs (Binary Large Objects) or in a file system as image files in any graphic format. The system is completely portable across different software/hardware platforms and supports both basic operations on images (scaling, cuts of rectangular area, filtering of fluorescence intensity, contrast enhancement, etc.) and subject domain oriented operations (generation of a multiple-stained image from several single-stained ones, masking by a nuclear mask, background removal, image registration). The system architecture provides for a user interface based on standard Web browsers and the HTTP protocol, and allows to use both FireWall and Proxy servers. The developed system is being successfully used for processing of images in the FlyEx database (Poustelnikova *et al.*, 2004).

Availability: http://urchin.spbcas.ru/flyex, http://flyex.ams.sunysb.edu/flyex

Introduction

The quantitative approach to the analysis of gene expression information allows to reveal fine details in gene regulation (Jaeger *et al.*, 2004). Quantitative gene expression data can be extracted from confocal images of gene expression patterns as these images have very high quality and resolution (Sharpe, Hecksher-Sorensen, 2001). Unfortunately relative few methods are currently available for extracting quantitative data from such images (OME, Myasnikova *et al.*, 2001) and linking this data to other biological information. Images are for the most part stored either in file systems or in relational databases, which do not have any built-in tools for image processing, while the image processing is performed by graphic packages installed on a local computer. This approach requires to download an image from a database or copy an image file to a local computer in order to process it, as well as to insert the processing results in the database or file system again. Thus the development of software for quantification and analysis of gene expression *in situ* is an important task for bioinformatics.

In this work we present a software system, which is designed for on-line processing of images stored in databases or graphic files. This system is portable across software/hardware platforms, supports image processing in the multiuser mode and publishing data in the Internet. It realizes both basic operations on images (scaling, cuts of rectangular area, filtering of fluorescence intensity, contrast enhancement, etc.) and subject domain oriented operations (generation of a multiple-stained image from several single-stained ones, masking by a nuclear mask, background removal, image registration). The architecture of the system provides for a user interface, which is based on standard Web browsers and the HTTP protocol, and allows to use both FireWall and Proxy servers.

The system designed is being applied to process the digital images of segmentation gene expression patterns. The determination of segments is the subject of intensive research over the last two decades and the quantification and analysis of segmentation gene expression is an essential part of the work of many researchers (Houchmandzadeh *et al.*, 2002; Wu *et al.*, 2001; Stathopoulos, Levine, 2002).

System and Methods

Methods for image processing

The basic operations for image processing are implemented by use of the JMagick package, which represents a Java interface to the ImageMagick package. These packages are publicly available (Yeo, 2003; Still, 2003). The subject domain oriented methods for image processing were developed by the authors.

Data normalization. Quantitative gene expression data is rescaled in order to get rid of distortions caused by the presence of background signal. The method for removal of background signal is based on the observation that the level of a given gene expression in a null mutant embryo for that gene is well fit by a very broad two dimensional paraboloid. The background paraboloid is automatically determined from the areas of wild type embryos in which a given gene is not expressed and used to remove background from the entire embryo.

Registration. To eliminate small individual differences quantitative gene expression data is subjected to registration. Two registration methods were developed. Both methods are based on the extraction of ground control points (GCPs). For GCPs the extrema of the expression pattern of the *eve* gene are used. The affine coordinate transformation is applied to make the corresponding GCPs in different images coincide as closely as possible. In one registration method (the spline or SpA method) GCPs are extracted by a quadratic spline approximation, while in the other method (FRDWT or wavelet) the fast dyadic redundant wavelet transform is used (Myasnikova *et al.*, 2001; Kozlov *et al.*, 2002).

System architecture

Figure presents the architecture of the system

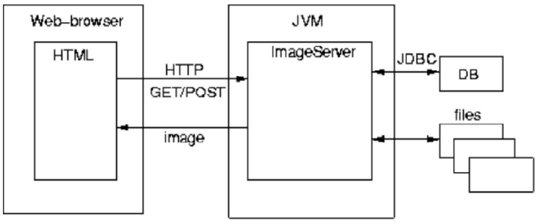


Fig. System architecture.

The architecture is three-tier, ImageServer (IS) is the middleware. IS is written in Java and represents an application server, which runs on the server under control of JVM (Java Virtual Machine). Use of the Java programming language provides platform independence of the system. IS is permanently waiting for client requests listening to an IP-port with a given number. Clients interact with IS via the HTTP protocol. IS realizes a subset of standard Web server functions, in particular, the GET and POST methods. Use of the HTTP protocol allows to use both FireWall and Proxy servers. To guaranty parallel and independent work of several clients a separate thread is created for each client's request. The images (operands), operations and other settings are specified as the parameters of the HTTP request. Clients are usually Web browsers, which call IS by including the standard tags <image> into the body of a HTML page. The <image> tags have to contain the server URL and the parameters. IS can access images which are stored in a relational database as BLOBs, image files in popular graphic formats (JPEG, GIF, TIFF, BMP, PNG etc.), as well as the files in RAW format, which represents the byte array of intensities for each image pixel. By default, the target image has the JPEG format, however the output format can be specified explicitly using the corresponding parameter. To provide software independence IS interacts with a database via the JDBC protocol.

Implementation and Discussion

Clients invoke IS by constructing the special URL containing parameters which identify the operands (i.e. images), the set of operations and their parameters. It is possible to perform the following operations on a single image: 1) scaling, the scaling coefficients are specified by a client; 2) cut of a rectangular area, the coordinates of the desired area are to be specified by a client; 3) filtering of fluorescence intensity, i.e. extraction of areas, where the expression level of at least one (or each) gene exceeds the predefined threshold or lies in the predefined interval, the threshold or interval of intensities is specified; 4) contrast enhancement; 5) background removal; 6) etc.

There are the following operations performed on a set of images: masking of one image by another, combination of up to three gray-scaled images into the color one, generation of an absolute value of difference between two images, registration of several images. It is possible to combine several operations in a single request.

The operation set described above allows to solve the following tasks: 1) superposition of a nuclear mask, i.e. extractions of segments (nuclei), in which a given gene is expressed; 2) generation of an image, which displays the expression patterns of all genes scanned in an embryo, each gene represented with its own color; 3) evaluation of the difference between two images; 4) estimation of the quality of quantitative data by examination of nuclear masks; 5) the accuracy of registration can be evaluated by comparing gene expression data before and after registration; 6) the level of background signal can be estimated from the image obtained as a result of subtraction of the background free image from the source one; 7) the variability of the expression of a given gene and changes of this variability in time or space can be estimated by subtraction of two images of the same age or by combining of up to three images.

There are several possible directions of further development of the system: increase of a number of basic and subject domain specific operations on images; the possibility to embed image processing operations into a SQL query; expansion of the system to a distributed multiagent system for on-line processing and analysis of images.

Acknowledgements

This work is supported by NIH grant R01 RR07801 and GAP award RBO-1286. We thank John Reinitz and David Kosman for key contributions to this work.

References

Houchmandzadeh B., Wieschaus E., Leibler S. Establishment of developmental precision and proportions in the early Drosophila embryo // Nature. 2002. 415(6873). P. 748–9.

Jaeger J., Blagov M., Kosman D., Kozlov K.N., Manu, Myasnikova E., Vanario-Alonso C.E., Samsonova M., Sharp D.H., Reinitz J. Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster* // Genetics. 2004. in press.

Kozlov K., Myasnikova E., Pisarev A., Samsonova M., Reinitz J. A method for two-dimensional registration and construction of the two-dimensional atlas of gene expression patterns *in situ* // In Silico Biol. 2002. V. 2. P. 125–141. http://www.bioinfo.de/isb/2002/02/0011.

Myasnikova E., Samsonova A., Kozlov K., Samsonova M., Reinitz J. Registration of the expression patterns of Drosophila segmentation genes by two independent methods // Bioinformatics. 2001. V. 17(1). P. 3–12.

Poustelnikova E., Pisarev A., Blagov M., Samsonova M., Reinitz J. A database for management of gene expression data *in situ* // Bioinformatics. 2004. in press.

Sharpe J., Hecksher-Sorensen J. 3D confocal reconstruction of gene expression in mouse // Mech. Dev. 2001. V. 100. P. 59–63.

Stathopoulos A., Levine M. Dorsal gradient networks in the Drosophila embryo // Developmental Biology. 2002. V. 246. P. 57–62.

Still M. Graphics from the command line // IBM developerWorks. 2003. http://www-106.ibm.com/developerworks/library/l-graf/?ca=dnt-428.

Swedlow J.R., Goldberg I., Brauner E., Sorger P.K. Informatics and quantitative analysis in biological images // Science. 2003. V. 300. P. 100–102.

Wu X., Vasisht V., Kosman D., Reinitz J., Small S. Thoracic patterning by the *Drosophila g*ap gene *hunchback* // Developmental Biol. 2001. V. 237. P. 79–92.

Yeo E. JMagick http://www.yeo.nu/jmagick/. 2003.

TOPICAL CLUSTERING OF BIOMEDICAL ABSTRACT BY SELF ORGANIZING MAPS

Fattore M., Arrigo P.*

CNR ISMAC, Section of Genoa,Via De Marini 6,16149 Genova, Italy
* Corresponding author: e-mail: arrigo@ge.ismac.cnr.it

Keywords: *text mining, conceptual clustering, self-organizing maps, information extraction*

Summary

Motivation: One of the major challenges in the post-genomic era is the speed up of the process of identification of molecular targets related to a specific pathology. Even if the experimental procedure have greatly enhanced the analytical capability, the textual data analysis still play a central role in the planning of the experiments or for database construction. The extraction of relevant information from the published paper requires a lot of time; tools that automatically cluster together the retrieved documents into topic categories labelled by specific relevant keywords can give a great support to this activity.

Results: In this paper we present the a application of document clustering system based on Self-Organizing maps to cluster PUBMED abstracts and for the extraction of class specific terms that allow to select the items that are related to some specific topics. The system allows the discrimination of different groups of items and gives an index of relevance for the terms. We have tested the system on a small test sample of PUBMED abstract related to the CDK5 proteins.

Availability: the software is available at the following site: http://www.biocomp.ge.ismac.cnr.it

Introduction

One of the major challenges in the post genome era is the screening of different data sources in order to perform integrated data mining. The biological literature is a major repository of knowledge. Many biological databases draw much of their content from a careful curation of this literature. However, as the volume of literature increases, the burden of data warehousing increases. Text mining may provide useful tools to support this activity. Literature mining is the process of extracting and combining facts from scientific publications. In recent years, many computer programs have been designed to extract various molecular biology findings from Medline abstracts or full-text articles [1, 2].

The textual knowledge discovery offers powerful methods to support knowledge discovery and the construction of topic maps and ontologies [3]. The challenge is to manage the increasing volume, complexity and specialization of knowledge expressed in this literature.

Although information retrieval or text searching is useful, it is not sufficient to find specific facts and relations. Information extraction methods are evolving to extract automatically specific, fine-grained terms corresponding to the names of entities referred to in the text, and the relationships that connect these terms. Many tools for text mining are founded on clustering methods [4] because the biomedical abstracts are not always organized into topic classes; these algorithms allow to partition the data set into clusters that can be separately analysed.

Method

The method proposed here is founded on SOM paradigm. The analysis is performed in different step. The different phase of the program are briefly sketched in the flowchart below.

The system perform two main activities: a) the linguistic data preprocessing, b) the data classification and mapping.

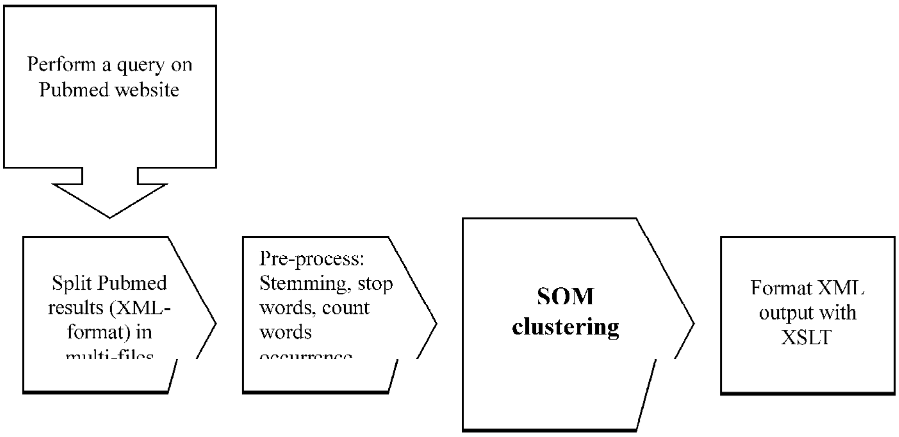

Perform a query on Pubmed website
Split Pubmed results (XML-format) in multi-files
Pre-process: Stemming, stop words, count words occurrence
SOM clustering
Format XML output with XSLT

Fig. 1. The flowchart of the document clustering method.

Data collection phase and linguistic preprocessing. The sample set is been retrieved from NCBI PUBMED server by using the standard query form. For our test application we have extracted a sample of 556 abstracts by using the keyword 'cdk5'. We have selected this protein because it seem to play a critical role in neurodegenerative diseases and in the number or published paper is not to large. Another reason is the relative influence of abstracts related to surgical or clinical researches; this constrain is relevant because we want to classify molecular biology topics. The retrieved document has been submitted to a linguistic preprocessing. First of all we have removed the 'stopwords' from the abstract text. This operation essentially cancels some irrelevant terms (articles, conjunctions, pronouns) that can affect the frequency count. After the stopwords removal the text is been subjected to stemming procedure [5]: the stemming performs the elimination of prefix and suffix. The set of stemmed tokens are submitted to the frequency count in each document. We take into account both the frequency of each term both the number of document in which the term is present. This phase generate a small 'dictionary' of terms (4952 tokens).

Document classification. In this phase we use the SOM for to cluster the sample. Before the training we need to convert the document set into a vectorial space representation. Given a set of documents **D**={d_1, …, d_n} each one is represented by a vector **v**={t_1, …, t_k} where t_k represent a word in the dictionary. Each term is represented in a numerical way as TFIDF (Term frequency, inverse document frequency); this parameter is computed in the following way:

$$\mathrm{TFIDF}=f(\mathrm{term})*[\log(N/nf)],$$

where ***f(term)*** is the occurrence frequency of the term in a document, ***N*** is the number of the document in the training set and ***nf*** is the subset of D that include the token term.

The conversion procedure originates a training set of d-dimensional vectors that can be processed by the network. The tokens are been ranked according their frequency of occurrence in the overall document set; each vector location represents the token and its rank. In order to reduce the vector dimensionality for this application we have applied an heuristic rule founded on the fraction (#_doc) of documents that include a specific term:

$$[|D|/10] < \#\_doc < [|D| - |D|/4],$$

where |D| is the cardinality of the document set.

The Self-organizing map is an unsupervised method for classification, the system use a set of computational elements arrayed in a two dimensional way. For the present application we have used the canonical version of the Kohonen algorithm [6]. At the end of the learning phase we have extracted the relevant term according the ordering properties of the SOM. We have ranked the final weights vectors in a descending order and then we have considered the terms associated to the highest values.

Implementation

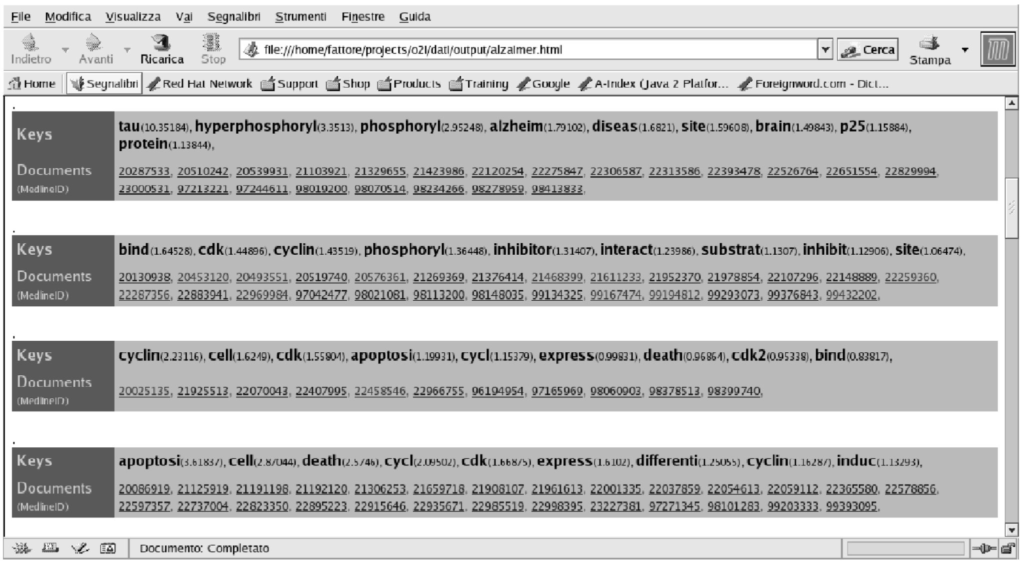
The software has been implemented in JAVA and run under LINUX RED HAT version 9.0. The system receives as input XML PUBMED abstracts. The output show the abstracts assigned to each node. For this application we have used a fixed rectangular lattice (squared) of 10 x 10 nodes. The current output is in a HTML format. The HTML output allows the user to click on the MEDLINE_ID code to view the requested abstract. The complete static map is available at the following site: http://www.biocomp.ge.ismac.cnr.it

Results and Discussion

The proposed system generate a map of grouped documents, each cluster is characterised by a set of the high relevance terms. The figure below shows an output sample of the topic map obtained by using the test set. This figure shows a partial visualization of the document map. For each node there is the vector of the representative terms and the documents. For instance, in the output sample below, the first and second node can be described by the following representative vectors:

N1→{tau∧hyperphosphorilation∧and phoshorilation∧alzheimer∧…}

N2→{bind∧cdk∧cyclin∧phosphorilation∧inhibition…..}



Keys	**tau**(10.35184), **hyperphosphoryl**(3.3513), **phosphoryl**(2.95248), **alzheim**(1.79102), **diseas**(1.6821), **site**(1.59608), **brain**(1.49843), **p25**(1.15884), **protein**(1.13844),
Documents (MedlineID)	20287533, 20510242, 20539931, 21103921, 21329655, 21423986, 22120254, 22275847, 22306587, 22313586, 22393478, 22526764, 22651554, 22829994, 23000531, 97213221, 97244611, 98019200, 98070514, 98234266, 98278959, 98413833,
Keys	**bind**(1.64528), **cdk**(1.44896), **cyclin**(1.43519), **phosphoryl**(1.36448), **inhibitor**(1.31407), **interact**(1.23986), **substrat**(1.1307), **inhibit**(1.12906), **site**(1.06474),
Documents (MedlineID)	20130938, 20453120, 20493551, 20519740, 20576361, 21269369, 21376414, 21468399, 21611233, 21952370, 21978854, 22107296, 22148889, 22259360, 22287356, 22883941, 22969984, 97042477, 98021081, 98113200, 98148035, 99134325, 99167474, 99194812, 99293073, 99376843, 99432202,
Keys	**cyclin**(2.23116), **cell**(1.6249), **cdk**(1.55804), **apoptosi**(1.19931), **cycl**(1.15379), **express**(0.99831), **death**(0.96864), **cdk2**(0.95338), **bind**(0.83817),
Documents (MedlineID)	20025135, 21925513, 22070043, 22407995, 22458546, 22966755, 96194954, 97165969, 98060903, 98378513, 98399740,
Keys	**apoptosi**(3.61837), **cell**(2.87044), **death**(2.5746), **cycl**(2.09502), **cdk**(1.66875), **express**(1.6102), **differenti**(1.25055), **cyclin**(1.16287), **induc**(1.13293),
Documents (MedlineID)	20086919, 21125919, 21191198, 21192120, 21306253, 21659718, 21908107, 21961613, 22001335, 22037859, 22054613, 22059112, 22365580, 22578856, 22597357, 22737004, 22823350, 22895223, 22915646, 22935671, 22985519, 22998395, 23227381, 97271345, 98101283, 99203333, 99393095,

Fig. 2. Example of output of cdk5 topical map: each cluster is represented by its relevant keywords. Below the keys are reported the PUBMED code of the abstract in the cluster. Each code is clickable by the user.

The output format allows the user to select, according the more relevant words, only a specified subset of abstracts. The capability to perform an optimal topical clustering is an essential step for the identification of the cluster specific words; only the characterization of the cluster tags allows applying a concept extraction procedure. Our aim is the development and implementation of an textual mining web services than can give to the user documentation conceptually homogeneous.

References

1. Nahm Y., Mooney R.J. Text mining and Information extraction AAAI // Symposium on Mining Answer from Text and Knowledge Bases Stanford. 2002.
2. Nenadic G. *et al*. Terminology-driven literature and knowledge acquisition in biomedicine // Intl. J. of Medical Informatics. 2002. V. 67. P. 33–48.
3. Baker P.G. *et al*. An ontology for bioinformatics applications // Bioinformatics. 1999. V. 15(6). P. 510–20.
4. Chen J.N., Chang J.S. Topical clustering of MRD sense based on information retrieval techniques // Computational Linguistics. 1998. V. 24(1). P. 61–95.
5. Porter M. An algorithm for suffix stripping // Program. 1980. V. 14, N 3. P. 130–137.
6. Kohonen T. Self-Organizing maps. Springer-Verlag, 2001.

GIBBS SAMPLER FOR IDENTIFICATION OF SYMMETRICALLY STRUCTURED, SPACED DNA MOTIFS WITH IMPROVED ESTIMATION OF THE SIGNAL LENGTH AND ITS VALIDATION ON THE ArcA BINDING SITES

Favorov A.V.* [1], ***Gelfand M.S.*** [1, 2, 3], ***Gerasimova A.V.*** [1], ***Mironov A.A.*** [1, 3], ***Makeev V.J.*** [1, 4]

[1] State Scientific Centre "GosNIIGenetica", 1st Dorozhny pr., Moscow, 117545, Russia; [2] Institute for Problems of Information Transmission, Russian Academy of Sciences, Bolshoi Karetny per. 19, Moscow 127994, Russia; [3] Dept of Bioengineering and Bioinformatics, Moscow State University, Lab. Bldg B, Vorobiovy Gory 1-37, Moscow 119992, Russia; [4] Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Vavilova 32, Moscow 119991, Russia
* Corresponding author: e-mail: favorov@sensi.org

Keywords: *DNA motifs detection, weight matrix, spaced structured motif, palindrome, tandem repeat, Gibbs sampler, Kullbak entropy distance, ArcA, respiratory regulation*

Resume

Motivation: Extraction of a common motif from a set of unaligned sequences is often applied to identify DNA sites that are recognized by regulatory transcription factors. The factors commonly recognize symmetrically structured DNA motifs, either inverted or direct repeats. The binding sites are often spaced. Every sequence fragment may carry no binding site. The motif length is usually unknown.

Result: We present a modification of the Gibbs sampling motif extraction algorithm, SeSiMCMC (Sequence Similarities by Markov Chain Monte Carlo), which finds structured motifs of symmetric types, as well as motifs without any explicit symmetry, in a set of unaligned DNA sequences. It employs an improved motif length and gap length estimators and accurately takes into account sequences that do not contain any motif. We have applied the algorithm to a set of upstream regions of the *E. coli* genes, which are known to be regulated by ArcA. Using comparative genomics techniques on the basis of the identified motif, we have found 23 genes in the *E. coli*. Fourteen of these are mentioned in literature as involved in a respiratory regulon.

Availability: The WWW interface of the program, its FreeBSD (4.0) and Windows 32 console executables and documentation are available at http://bioinform.genetika.ru/SeSiMCMC .

Introduction

Extraction of a common motif from a set of unalignment sequence fragments (also known as the multiple local alignment problem, MLA) is often applied to identify DNA sites that are recognized by transcription regulatory factors. This approach is based on the assumption that functionally related DNA segments contain similar nucleotide subsequences.

Usually, the analysis starts from a sample of DNA sequences, the majority are supposed to contain a protein-binding site (or another characteristic segment). These segments carry the same signal and therefore are instances of the same motif. The objective is to classify all DNA sequence data into motif instances and the remaining background in an optimum way. The motif is represented by a positional weight matrix (PWM) (Berg, von Hippel, 1987; Lawrence *et al*., 1993); the one background is the modelled by independent letters with fixed probabilities. Thus the MLA problem is reformulated as maximization of the posterior of the foreground-background partition given the sequence.

Here we present a modification (Favorov *et al*., 2002) strongly suggested by extensive practice of analysis and prediction of gene co-regulation in prokaryotes (e.g., Gelfand *et al*., 2000; Thompson, Rouchka *et al*., 2003). Particularly, a signal recognized by a prokaryotic transcription factor often exhibits a structure of an inverted or direct repeat. Therefore, in extention of (Favorov *et al*., 2002), we included user-defined symmetry in the probabilistic motif model and a possiblity for a motif, either symmetric or not, to be spaced. Such a motif contains unimportant positions in its middle. One usually does not know in advance the motif length as well as the spacer length. Thus, the program estimates the optimum values for these two lengths during motif detection. The training set may erroneously contain biologically irrelevant sequences, which do not contain the target motif site. To account for this possibility, we explicitly enhanced the core probabilistic model with the expectation of a motif absence in a sequence.

We intended to create a specialized tool for finding weak structured motifs with spacers of unknown length. To this end, we designed a probabilistic model and an optimization procedure as modifications of the classical algorithm of Lawrence *et al*., 1993. We suppose that the specialized tool can be more adequate for this particular task than the universal.

Algorithm and Implementation

Our modification of the algorithm allows a user to specify the motif as containing two boxes either as a direct repeat or as a palindrome, possibly with a space of unknown length in the middle. If a symmetry is specified, the foreground model reflects it. The core procedure for selection of the best (or almost the best) set of sites is as follows. We organise a cycle of one-by-one site positions updates. At each step, we select only one sequence. The site absence is treated as a position of a specific type ("null"). At each step, we collect the nucleotide statistics for the internal site positions and for the background from all sequences except the one updated at this step. The Bayesian posterior distribution is obtained for a site position in the current sequence; we draw the new site position (including the "null") from this distribution. The process is repeated cyclically until the chain of site sets converges (i.e. the step-to-step changes become small). The algorithm is similar to that described in (Lawrence *et al*., 1993), with the difference that we process the possibility of a site absence in the Bayesian way at every updating step.

In fact, the algorithm optimises the self-consistence of a set of site positions, so it is very sensitive to changes in the mutual location of the sites, but it is quite tolerant to all-as-one shifts on the site position set. To solve this problem, we adjust the results from time to time after the core algorithm converges to a sufficient degree, then restart the core. The adjustment is a deterministic search for the best solution among all the possible cooperative shifts of the local alignment of sites. As a quality function we maximize the motif information content per letter (ICL). The procedure is similar to that described in (Lawrence *et al*., 1993) but differs from it by latter in a number of aspects. The information content calculation has an improved spatial component. The site length is evaluated at the adjustment stop. Moreover, the same procedure is used to determine whether the motif is spaced. We assume that some middle columns of MLA may be not correlated, and in this case the motif has a symmetric spacer, which corresponds to the background probabilistic model. Thus, we extend the spacer for every cooperative shift of sites during the adjustment until we obtain the local maximum of the ICL for that shift.

The SeSiMCMC software is written as C++ (gcc 3.x). Executable files for FreeBSD and Windows 32 console and the program documentation describing the command line and the configuration file control interfaces are available at the project site http://bioinform.genetica.ru/SeSiMCMC. Also, the site provides a web interface for the program with input forms.

The result, switch ArcA of the type of respiration

In *Escherichia coli*, gene expression is dependent on redox conditions, which is partly mediated by the Arc signal transduction system. The phosphorylated form of ArcA protein (ArcA-P) represses certain target operons (e.g. *icd*, *lld*, *sdh* and *sodA*) or activates others (e.g. *cyd* and *pfl*) by interacting with promoter DNA. We used the tool to search for a common motif in upstream regions of the genes, which were extracted as ArcA-regulated (Lynch, Lin, 1996) from the DPInteract database (http://arep.med.harvard.edu/dpinteract/). The parameters for the motif search were selected as a possibly spaced direct repeat of length between 6 and 22 bases located at any DNA strand. As a result (Favorov, Gerasimova, 2003), a 15-nucleotide motif (Fig.) was obtained, which refines the known ArcA binding site structure (McGuire *et al.*, 1999).

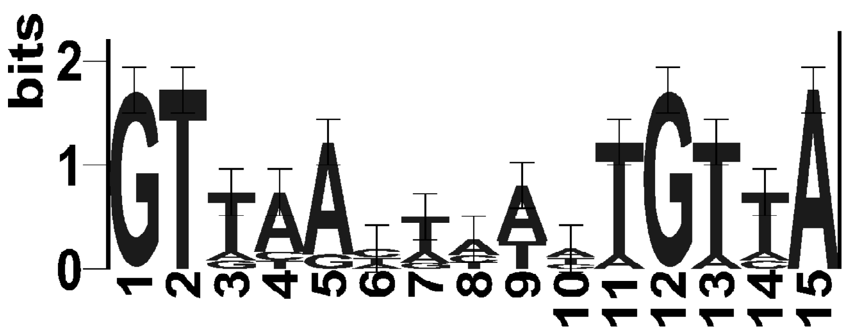


Fig. ArcA binding site motif logo, according to (Favorov, Gerasimova, 2003). Created by WebLogo (http://www.bio.cam.ac.uk/cgi-bin/seqlogo/logo.cgi).

The found set of sites was used to create a PWM. Genome Explorer software (Mironov *et al.*, 2000) was used in all studies. A combination of the matrix matches with genome comparison allowed us to identify 23 *E. coli* genes with the upstream ArcA box scored higher than 4,25 and with at least two orthologs in *Y. pestis, P. multocida* and *V. vulnificus*, which carry an ArcA boxes scored at least 4,00 in the upstream. One of the found genes is the ArcA protein gene itself. Fourteen of these genes are referred to the literature as oxygen-dependently regulated (Gerasimova *et al.*, 2004).

The probability of the null-hypothesis of random gene selection by the recognition rule can be given the high estimate of 500 oxygen-dependent genes among 4,404 genes in the full *E. coli* genome. Fisher's test for the "14 9 // 500 3904" four-pole table gives the null-hypothesis probability of about 2x10^{7}. Thus, the null-hypothesis can be reliably rejected.

Discussion

All in all, SeSiMCMC is a tool for multiple local alignment of a set of DNA sequence fragments that is based on a modification of the Gibbs Sampling (Lawrence *et al.*, 1993) algorithm. Our primary objective was to create a computationally efficient tool that employs user-defined motif symmetry and evaluates the motif length from the data. Sequence fragments in the initial set can have an arbitrary orientation, and there is a probability for a sequence to contain no sites. SeSiMCMC was tested on several sets of bacterial regulatory regions, where we were able to extract regulatory motifs without supplying the motif length and sequence orientation (see the project's Web site).

A recognition rule created using the SeSiMCMC algorithm revealed 23 ArcA-regulated genes in *E. coli,* fourteen of them are referred to in the literature as oxygen-dependently regulated.

Acknowledgements

We are grateful to D. Rodionov for useful discussion and to L. Danilova for assistance with the data. This study was partly supported by grants from the Howard Hughes Medical Institute (55000309 to M. Gelfand), the Ludwig Cancer Research Institute (CRDF RBO-1268-MO-02 to M. Gelfand), the Russian Fund of Basic Research (04-07-90270 to V. Makeev) and from the Program in Molecular and Cellular Biology of Russian Academy of Sciences (to V.G. Tumanian).

References

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

Favorov A.V., Gelfand M.S., Mironov A.A., Makeev V.J. Yet aother digging for DNA motifs Gibbs sampler // BGRS'2002. Proceedings. 2002. V. 1. P. 31–33.

Favorov A.V., Gerasimova A.V. Yet another digging-for-DNA-motifs Gibbs sampler. MCCMB'03. Proceedings. 2003. P. 67–69.

Gerasimova A.V., Gelfand M.S., Makeev V.J., Mironov A.A., Favorov A.V. Primary description of ArcA regulon in gamma-proteobacteria genomes on the base of the regulatory protein binding site computational recognition // Biophysics. 2004. Moscow, in press.

Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., Wootton J.C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment // Science. 1993. V. 262. P. 208–214.

Lynch A.S., Lin E.C.C. Regulation of Gene Expression in *E. coli* / Eds. A.S. Lynch, E.C.C. Lin. 1996. P. 361–376. (Austin, TX: Landes).

McGuire A.M., De Wulf P., Church G.M., Lin E.C.C. A weight matrix for binding recognition by the redox-response regulator ArcA-P of *Escherichia coli* // Mol. Microbiol. 1999. V. 32 (1). P. 219–221.

Mironov A.A., Vinokurova N.P., Gelfand M.S. Software for analysis of bacterial genomes // Mol. Biol. 2000. V. 34(2). P. 222–231.

Thompson W., Rouchka E.C., Lawrence C.E. Gibbs recursive sampler: finding transcription factor binding sites // Nucleic Acids Res. 2003. V. 31(13). P. 3580–3585.

MATHEMATICAL MODELLING AND ANALYSIS OF THE FIXATION PROCESS OF DISCRETE GENETIC STRUCTURES IN A MENDELIAN ONE-LOCUS POPULATION OF DIPLOID ORGANISMS

Frisman E.Ya.*[1]**, Zhdanova O.L.**[2]

[1] Complex Analysis of Regional Problems Institute FEB RAS, Birobidzhan, Russia, e-mail: frisman@mail.ru; [2] Institute of Automation and Control Processes FEB RAS, Vladivostok, Russia
* Corresponding author: e-mail: axanka@iacp.dvo.ru

Keywords: *mathematical model, computer analysis, adaptation, genetic distributions, mutations*

Summary

Motivation: An explaining the mechanisms of formation of discrete biological taxa is one of the main problems of evolutionary theory.

Results: An integral model of the evolution of a Mendelian one-locus population of diploid organisms with continual allele diversity developing under density-limiting conditions or without density limitation has been proposed and analyzed. The model was used to study the mechanism of the appearance of discrete genetic structures, i.e., the fixation of a limited number of alleles. Local resistance of the resultant genetic distributions to homogeneous equiprobable mutations has been demonstrated.

Introduction

To explain the mechanisms of formation of discrete biological taxa is one of the main problems of evolutionary theory. The genetic diversity within a species is often discrete and strictly limited. The reason is hardly the discreteness of the "heredity carrier" itself, i.e., DNA, consisting of monomers. A protein consists of several hundred aminoacids. Mutational variation may yield a vast diversity of molecules of a given protein, with most of these molecules functioning normally (Altukhov, 2003). However, only one form of a given protein is usually fixed in the populations. Two forms of a protein are seldom fixed; three forms, even more seldom; etc. What is the mechanism of fixation of some alleles and loss of others? There are two main hypothesis answering this question: (1) random loss of alleles because of gene drift and (2) balanced polymorphism determined by the selective advantage of heterozygotes. Both of them have supporters and opponents (Crow, Kimura, 1971; Lewontin, 1974; Altukhov, 2003), but neither provides a definite solution to the problem. Our study is one more attempt to analyze this issue.

Model

Let us consider a large Mendelian panmictic sexless population in which the inheritance of a certain character is determined by one with gene infinite number alleles. The genetic-structure and population dynamics in this population can be described by the following system of equations:

$$\begin{cases} x_{n+1} = \overline{W}_n x_n \\ q_{n+1}(\tau) = q_n(\tau)\left(\int_0^1 W(\xi,\tau) q_n(\xi) d\xi\right) / \overline{W}_n \\ \overline{W}_n = \int_0^1\int_0^1 W(\xi,\tau) q_n(\xi) q_n(\tau) d\xi d\tau \end{cases} . \quad (1)$$

Here, n is the ordinal number of the generation, x_n is the population number, the function $q_n(\tau)$ is the frequency density of allele τ in the population in the n-th generation, W(ξ,τ) is the adaptation function of the genotype (ξ,τ), $\overline{W}_n$ – is the population mean adaptation in the n-th generation, each of ξ and τ being the allele markers and may be any real number within the interval [0, 1]. In fact, we approximate a finite-dimensional situation by an infinite-dimensional one (Gorban, Khlebopros, 1988) in order to use all the possibilities of continuous functions analysis.

If density-dependent selection takes place in the population, then fitnesses are decreasing functions of the population size. The exponential dependence of fitness on population size is suitable for analysis. This dependence can be written as follows:

$$W(\xi,\tau,x_n) = \exp(R(\xi,\tau)(1 - x_n / K(\xi,\tau))) . \quad (2)$$

In this case, each genotype is characterized by two parameters, R(ξ,τ) and K(ξ,τ) (the Malthusian and the resource parameters, respectively) (Evdokimov, 1999).

The special case of the integral dynamic model of a one-locus diallelic population with adaptations independent of the population number, i.e., in the absence of density control, is of special interest. This model is a logical generalization of the classical model of the dynamics of a one-locus diallelic population with constant adaptations of genotypes (Ratner, 1977; Frisman, 1986) extending it to the case when there is a continual number of alleles of one locus.

In the absence of density limitation, the population number dynamics is of no interest, because it will either infinitely grow (if $\overline{W}_n$ > 1) or constantly decrease (if $\overline{W}_n$ < 1). Therefore, let us consider separately the dynamics of allele-frequency density in case of unlimited population:

$$q_{n+1}(\tau) = q_n(\tau)\left(\int_0^1 W(\xi,\tau) q_n(\xi) d\xi\right) / \int_0^1\int_0^1 W(\xi,\tau) q_n(\xi) q_n(\tau) d\xi d\tau . \quad (3)$$

Here the adaptation function W(ξ,τ) is also a function of the Malthusian and the resource parameters of (ξ,τ)–genotype.

Then we studded the effect of mutations on the dynamics of the integral model of an unlimited population. Let mutations occur before selection; then, the allele-frequency density in the n-th generation after mutation takes the form

$$\tilde{q}_n(\tau) = \int_0^1 q_n(\xi) \mu(\xi,\tau) d\xi, \quad (4)$$

where $\mu(\xi,\tau)$ is the probability density of the mutation from ξ to τ.

After this, selection takes place:

$$q_{n+1}(\tau) = \tilde{q}_n(\tau)\left(\int_0^1 W(\xi,\tau)\tilde{q}_n(\xi) d\xi\right) / \int_0^1\int_0^1 W(\xi,\tau)\tilde{q}_n(\xi)\tilde{q}_n(\tau) d\xi d\tau . \quad (5)$$

Stationary solutions of Eq. (5) are continuous functions explicitly depending on mutations described by the function $\mu(\xi,\tau)$. It may be expected that, if selection “aims” at creating a discrete distribution of allele “frequencies” in the population, mutations will “smear” the almost discrete peaks.

We assume also that mutations are homogeneous and equiprobable (6) and then Eq. (6) may be rewritten as (7).

$$\mu(\xi,\tau)=\begin{cases}\delta, \forall\xi\neq\tau\\ 1-\delta, \forall\xi=\tau\end{cases},\ \xi,\tau\in[0,1]. \tag{6}$$

$$\widetilde{q}_n(\tau)=(1-\delta)q_n(\tau)+\delta. \tag{7}$$

Results

It has been analytically demonstrated that both of the evolution described by system of equations (1) and those described by equation (3) transforms correct genetic distributions into correct ones (Frisman, Zhdanova, 2003). Furthermore, we have found that, if there is fitness heterogeneity in the space of genotypic classes (in mathematical terms, this means that $\{\int_0^1 W(\xi,\tau,x_n)d\xi=\Psi(\tau,x_n)\neq f(x_n)$, where f($x_n$) is the function of population number alone and does not depend on τ – in case of density-dependent natural selection} and $\{\int_0^1 W(\xi,\tau)d\xi\neq\text{const}$ – in case of unlimited population}), then the proposed model has no continuous, stationary distributions of allele frequencies (Frisman, Zhdanova, 2003). Therefore, it can be expected that the evolution of (1) (and (3)) will result in the transformation of continuous density distributions of allele frequencies into vastly inhomogeneous ones provided that there is diversity of fitness values.

The dynamics of the proposed model was studied numerically. We have analyzed the population number dynamics and changes in population genetic structure with time for different variants of the initial distributions of allele "frequencies" and adaptation functions. The stationary genetic distributions being vastly inhomogeneous were obtained really in case of adaptations heterogeneity presence. Moreover, the stationary genetic distributions obtained have a few peaks number.

Then we have performed numerical analysis of the dynamic behavior of the system with mutations of (7, 5) at a fixed value $\delta = 0.1$. It has been shown that mutations can slightly smear the "discrete" distribution and, in addition, increase the size of the peak and considerably change the distribution pattern. Note, however, that markedly heterogeneous distributions with small numbers of peaks are still observed, the mutation rate that we set in the model ($\delta = 0.1$) is substantially exaggerated compared to the actual value in natural populations (10^{-5}). Note that mutations have increased genetic diversity in some model cases.

Discussion

Thus, the results of our study demonstrate that, "in the general case," even an infinitely (continually) large diversity of alleles is reduced to a small number of discrete alleles in the course of evolution under strictly determined conditions. Even introduction of some equiprobable mutations into the model does not result in stable homogeneous distributions. The mutations process somewhat "smears" the resultant distributions; however, a few almost discrete peaks are preserved, although both their number and heights may increase. Apparently, this situation will not change even in the case of a large (but finite) number of original alleles. The dynamic equations are such that evolution does not lead to homogeneous distributions of large numbers of alleles. Typically, distributions with small numbers of forms appear. Apparently, this explains why the allelic diversities of many genes are substantially limited in natural populations.

Acknowledgements

This study was supported by the Russian Foundation for Basic Research (project Nos. 03-01-00044-a and 04-04-97000-p).

References

Altukhov Yu.P. Genetic processes in populations. Akademkniga, Moscow, 2003.

Crow J., Kimura M. An introduction to population genetics theory // Princeton Univ. Press, Princeton, New Jersey, 1971.

Gorban A.N., Khlebopros R.G. Darwin's demon: the idea of optimality and natural selection. Nauka, Moscow, 1988.

Evdokimov E.V. Problems of regular behavior and determined chaos in the main models of population dynamics: theory and experiment. Doctoral (Biol.) Dissertation. Krasnoyarsk, 1999.

Frisman E.Ya. Primary genetic divergence (Theoretical analysis and modeling) // Dal'nauka. Tsentr Akad. Nauk SSSR. Vladivostok, 1986.

Frisman E.Ya., Zhdanova O.L. An integral model of the dynamics of size and genetic composition of a mendelian single-locus population of diploid organisms // Tr. Dalnevost. Gos. T. Univ. 2003. V. 133. P. 157–164.

Lewontin R.C. The genetic basis of evolutionary change // Columbia Univ. Press, New York, 1974.

Ratner V.A. Mathematical population genetics (An elementary course). Novosibirsk: Nauka, 1977.

SEVEN CLUSTERS AND UNSUPERVISED GENE PREDICTION

Gorban A.N.[1,2], *Popova T.G.**[1], *Zinovyev A.Yu.*[3]

[1] Institute of Computational Modeling SB RAS, Krasnoyarsk, Russia; [2] Institute of polymer physics, ETH, Zurich, Switzerland; [3] Institut des Hautes Études Scientifiques, Bures-sur-Yvette, France
* Corresponding author: e-mail: tanya@icm.krasn.ru

Keywords: *triplet frequency, visualization, gene recognition, unsupervised learning*

Summary

Motivation: The effectiveness of most unsupervised gene-detection algorithms follows from a cluster structure in oligomer distributions. Existence of this structure is implicitly known but it was never visualized and studied in terms of data exploration strategies. Visual representation of the structure allows deeper understanding of its properties and can serve to display and analyze characteristics of existing gene-finders.

Results: The cluster structure of genome fragments distribution in the space of their triplet frequencies was revealed by pure data exploration strategy. Several complete genomic sequences were analyzed, using visualization of distribution of 64-dimensional vectors of triplet frequencies in a sliding window. The structure of distribution was found to consist of seven clusters, corresponding to protein-coding genome fragments in three possible phases in each of the two complementary strands and to the non-coding regions with high accuracy. The self-training technique for automated gene recognition both in entire genomes and in unassembled ones is proposed.

Availability: http://www.ihes.fr/~zinovyev/bullet/

Introduction

In Fig. 1a, b one can see two projections of the 3D data scatters. Each point represents a genome fragment clipped by the sliding window and presented by its non-overlapping triplet frequencies. One see the seven cluster structure of the distribution. The central cluster (Fig. 1a, c) corresponds to non-coding genome fragments while side clusters correspond to protein-coding fragments related to three possible phases (reading frames at translation) in each of the two complementary strands with higher than 90 % accuracy on nucleotide level.

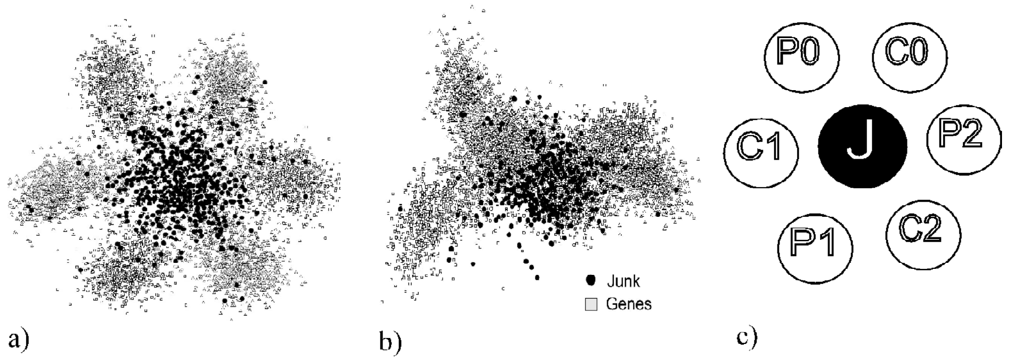


Fig. 1. Visualisation of *C.crescentus* (GenBank NC_002696) genome fragments distribution (a, b) and general seven clusters structure (c).

Seven clusters structure of genome fragments distribution plays important role in ability of modern gene-finders for unsupervised (and, to lesser extent, also for supervised) learning in prokaryotic genomes (Audic *et al.*, 1998; Baldi, 2000). Actually existence of the structure makes the prokaryotic gene-finding so efficient. While using seven hidden states for hidden Markov model in gene-prediction was introduced long ago (see, for example, Borodovsky *et al.*, 1993) and being widely exploited so far, this structure was never visually presented and analyzed by pure data exploratory strategy.

Here we introduce (1) the seven cluster structure of genome fragments distribution in the space of non-overlapping triplet frequencies and as illustration a simple unsupervised procedure for detecting coding regions; (2) some features of coding regions and gene-finders that become evident after the seven clusters structure was revealed.

Model

Seven clusters structure for compact genomes. Consider we have DNA sequence – genome fragments or complete genome. It is converted into the set of 64-dimensional vectors of triplet frequencies as follows.

A sliding window of the length W (W is to be about average exon size – 200–400 in our studies) and centered at position i is characterized by non-overlapping triplet frequencies calculated throughout the window: starting from the first nucleotide and up to the end. So, each data point vector $X_i=\{x_{is}\}$ corresponds to i-th window and has 64 coordinates which are frequencies of all possible triplets s = 1,…,64.

The standard centering and normalization on unit dispersion procedure is then applied, i.e., $\tilde{x}_{is} = \frac{x_{is} - m_s}{\sigma_s}$, where m_s and σ_s is the mean value and standard deviation of the s-th triplet frequency in the dataset.

Visualization of this 64-dimensional dataset in projection onto the 3-dimensional linear manifold spanned by the first three principal vectors of the distribution gives the well-detected seven clusters structure (Fig. 1).

Analysis of the distribution shows that the central cluster (Fig. 1a) is formed by the points X_i taken from the non-coding genome regions while side clusters are formed by the points of protein-coding regions. More specifically (see Fig. 1c), cluster P0 corresponds to the case when countered triplets are codons of the direct strand genes, C0 – codons of complementary strand genes, but in complementary translation and read from back to front ("shadow" genes, because only direct strand is considered), clusters P1, P2, C1, C2 contain points from coding regions too but read with 1 and 2 nucleotides shift relative to gene start.

Simple unsupervised procedure for detecting coding regions. Scanning the sequence with the sliding window step divisible by three and applying some clustering algorithm (K-Means clustering in the 64-dimensional space in our case) one obtains homogeneous with respect to the cluster label regions within the sequence. The J cluster gives non-coding regions but other clusters mark out protein coding regions. In more detail the gene predicting algorithm can be found in (Gorban *et al.*, 2003).

To evaluate the ability of the procedure to detect genes the base-level sensitivity and specificity were calculated, which are commonly used in this case:

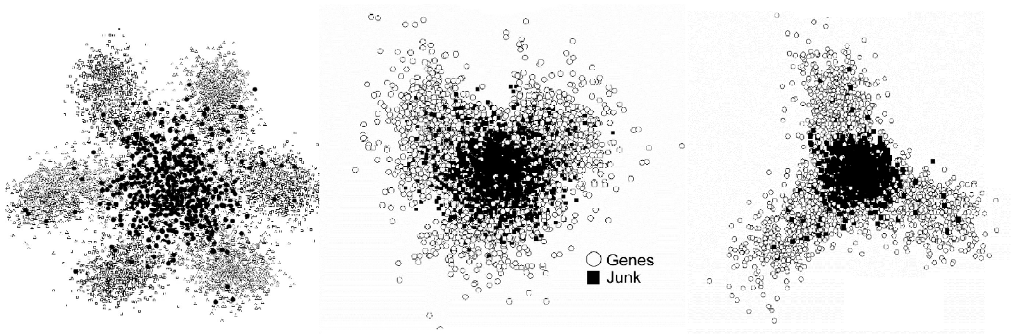
$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TP}{TP + FP},$$

where TP (true positives) is the number of coding bases predicted to be coding; FP (false positives) is the number of non-coding bases predicted to be coding, and FN (false negatives) is the number of coding bases predicted to be non-coding.

Results and Discussion

In this section we consider some results and observations obtained while analysing the seven cluster structure of some genomes and comparing it to well known facts and gene-finders.

The seven cluster structure and estimation of gene prediction accuracy. The cluster structure for some genomes is presented in Figure 2. The distribution in Fig. 2a shows very clear separation on seven clusters; no surprise that in this case unsupervised gene-prediction gives both high specificity and sensitivity.



a) Sn = 0.93; Sp = 0.97 b) Sn = 0.89; Sp = 0.91 c) Sn = 0.82; Sp = 0.93

Fig. 2. a) *C. crescentus* (GenBank NC_002696); b) *S. cerevisiae* chr.IV (GenBank NC_001136); c) *P. wickerhamii* (GenBank NC_001613).

The distribution of triples in *S. cerevisiae*, chr. IV (Fig. 2b) forms seven clusters as well; though they are not clearly seen on 2D-pictures, because two "phase triangles" P0-P1-P2 and C0-C1-C2, projected into the principal subspace are positioned on two parallel planes, perpendicular to the projection plane. Nevertheless simple clustering algorithm yields good prediction. The situation is worse in case of *P. wickerhhamii* mitochondrion genome. In this case distributions of triplets in the direct and reverse strands indeed overlap: P0 cluster overlaps with C1 cluster and so on. One can predict in this case that gene recognition procedures will often mix genes in the direct and reverse strands, though ORF-strategies can probably resolve this conflict.

So, visualization of datasets is useful to evaluate how reliable gene prediction could be.

Gene identification accuracy of our CLUSTER method and GLIMMER gene-finder. Choosing GLIMMER (version 2.02) (Salzberg *et al.*, 1998; Delcher *et al.*, 1999) for comparison was dictated by our desire to take a gene-predictor that uses no additional learning information, except one that can be extracted from the genetic sequence itself.

Table.

Genome	CLUSTER		GLIMMER	
	Sn	Sp	Sn	Sp
Helicobacter pylori	0.94	**0.95**	**0.96**	0.78
Haemophilus influenza	0.93	**0.88**	**0.96**	0.84
Escherichia coli	0.91	**0.87**	**0.96**	0.76
Bacillus subtilis	0.89	**0.95**	**0.97**	0.79
Caulobacter crescentus	0.89	**0.76**	**0.94**	0.60

The results of this comparison are shown in the Table. As one can see, the sensitivity of our method is lower in all cases, on the other hand specificity of our method is significantly better.

We have analysed why the GLIMMER produced a lot of false-positive predictions using visualization tool and seven cluster structure. Our analysis shows that 80 % of false positives for *C. crescentus* in the 64-dimensional space of triplet frequencies are closer to the C0 centroid, while only 2 % of true-positive predictions for *C. crescentus* are close to the C0-centroid. Such discrepancy cannot be explained simply by “presence of unknown genes” but it is due to some effect of this HMM-based predictor, which often takes “shadow” genes as positive predictions.

We have analysed why our clustering method produced more false-negatives than GLIMMER did for *E. coli* genome. It became clear from genome annotation: a half of missed genes are noted as predicted only by computational methods, other significant groups are ribosomal genes and transposases. It is known that ribosomal genes, some other highly expressed genes as well as horizontally transfered genes can have rather different (with respect to the average) codon usage, that is why our simple procedure based on triplet frequencies failed to predict them.

Some observations and conclusions

1. Our study shows relatively high performance of using only triplets for gene prediction in compact genomes. Using hexamer frequencies (that is common practice in modern gene-finders) can be more sensitive, but also can lead to certain undesirable “overfitting” effects and worse specificity.
2. The structure of codon usage over all genes in a genome is known to be inhomogeneous, especially in fast-growing organisms as *E. coli* and *B. subtilis*, however, the cluster structure shows a less within group dispersion than between groups dispersion. Thus the gene-prediction is possible even without preliminary genes classification.
3. Frequency normalization plays an important role in cluster structure formation. It indicates the importance in distinguishing coding and non-coding regions of those codons that may not have high frequency values but considerably change their frequencies after reading frame shift.
4. The proposed procedure for detecting genes is fully automated and requires no prior learning on known genes. The method can be applied for the rough annotation of unassembled genomes, since it does not require preliminary extraction of ORFs.
5. The cluster structure may be very useful in solving the problem of choosing a “good” learning dataset as it is not very well defined yet (see, for example (Mathe *et al*., 2002)).

References

Audic S., Claverie J.-M. Self-identification of protein-coding regions in microbial genomes // Proc. Natl Acad. Sci. USA. 1998. V. 95. P. 10026–10031.

Baldi P. On the convergence of a clustering algorithm for protein-coding regions in microbial genomes // Bioinformatics. 2000.V. 16. P. 367–371.

Borodovsky M., McIninch J. GENMARK: parallel gene recognition for both DNA strands // Comp. Chem. 1993. V. 17. P. 123–133.

Gorban A.N., Zinovyev A.Yu., Popova T.G. Seven clusters in genomic triplet distributions // In Silico Biology. 2003. V. 3. 0039. </isb/2003/03/0039>

Salzberg S.L., Delcher A.L., Kasif S., White O. Microbial gene identification using interpolated Markov Models // Nuc. Acids Res. 1998. V. 26(2). P. 544–548.

Delcher A.L., Harmon D., Kasif S., White O., Salzberg, S.L. Improved microbial gene identification with GLIMMER // Nuc. Acids Res. 1999. V. 27(23). P. 4636–4641.

Mathe C., Sagot M.F., Schiex T., Rouze P. Current methods of gene prediction, their strengths and weaknesses // Nucleic Acids Res. 2002. V. 30(19). P. 4103–4117.

MEASURING THE DISSIMILARITY BETWEEN GENE AND SPECIES TREES, THE QUALITY OF A COG

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

Institute for Information Transmission Problems, Russian Academy of Sciences, 127994, Moscow, Bolshoi Karetnyi per., 19, Russia
* Corresponding author: e-mail: lyubetsk@iitp.ru

Keywords: *evolution, phylogenetic trees, mathematical model computer analysis*

Summary

Motivation: The availability of genome-scale sequence data from diverse taxa makes it possible to derive new hypotheses about ancient evolutionary events from comparative analysis of large gene sets. Important groundwork of this goal is to find good strategies for comparing COG trees with species trees, to estimate the quality of the COGs and corresponding the trees to compare evolutionary models underlying the reconstructions and, in particular, to integrate approaches allowing inferences about evolutionary scenarios and gene duplication-loss patterns.

Results: In this study we reconstruct selected details of the ancestral history of Archaea and Bacteria within the outlined framework.

Introduction

It is well known that phylogenetic trees derived from different protein families are often incongruent and essentially differ from each other and the species tree. This may be explained by poor choice of the evolutionary model and associated reconstruction biases, as well as by discrepacies between gene and organism evolutionary history due to speciation, gene duplication, gene loss and gene gain (horizontal gene transfer, genesis from non-coding DNA), and so on. In this paper we consider several integral characteristics measuring dissimilarity between gene and species trees as estimates of the quality of the COG or the COG tree. We identify COGs (clusters of orthologous groups of proteins) of different quality. The main purpose of our analysis is reconstruction of selected details of the ancestral history of Archaea and Bacteria. For this analysis, we use the combined gene duplication-loss model extended to incorporate some gene gain events. Any method of tree comparison is based on an evolutionary hypothesis and the corresponding mathematical model of evolution. We compare two such hypotheses.

Models and Algorithms

We employ mapping from a gene tree G into a species tree S introduced in Mirkin *et al.* (Mirkin *et al.*, 1995), and extensively used in V'yugin *et al.* (V'yugin *et al.*, 2003). Suppose that two binary trees are given, a gene tree G (for a fixed COG) and a species tree S (including all species present in this COG). The unique *tree mapping* $\alpha: G \to S$ was defined in (Mirkin *et al.*, 1995; V'yugin *et al.*, 2003). We consider two methods of gene and species tree comparison. **The first method** is based on comparison of the combinatorial structures of the trees G and S. We use the tree mapping a and compare the neighborhood O_g of the gene g in the gene tree G and its image (under α) in the species tree S. We assume that the gene g is "ambiguous" in position on the species tree if $\alpha(g)$ and $\alpha(O_g')$ are far apart in the species tree S (where O_g' is exactly the neighborhood O_g with the gene g omitted), the numerical characteristic R_g is a measure of this difference. High values of R_g reflect positional ambiguity of the gene g in the species tree (for details refer to (V'yugin *et al.*, 2003)). In this paper we define **the integral characteristics of the gene tree** $<R_g> = (1/m)\Sigma_g R_g$, i.e. the mathematical expectation of the R_g statistic over the corresponding COG, where m is the number of genes in the COG. **The second method of comparison** is based on the gene duplication-loss model. A measure of *dissimilarity* $c(G,S)$ of a gene tree G and a species tree S (which is the sum of one–side duplications and intermediate nodes [1, 2]) was introduced. Thus, for any COG tree G we calculate the cost $F=c(G,S)$ (for α: $G \rightarrow S$). Further, we reduce the gene tree G_g by gradually removing and replacing genes g from the gene tree G and calculate the cost F_g (for α: $G_g \rightarrow S$). The relative change in the costs of two tree mappings α is calculated with the formula $dF_g = (F_g\text{-}F)/F$. The corresponding **integral characteristics** of the COG is the expectation $<dF_g>$ of dF_g over all its genes g. When we analyse a set of COGs, we denote by $<<R_g>>$ and $<<dF_g>>$ the expectations of $<R_g>$ and $<dF_g>$, respectively, for all the COGs.

Results and Discussion

A set of maximum-likelihood trees constructed for COGs was analysed in (V'yugin *et al.*, 2003). For any statistic $f(g)$ we consider the corresponding p-value $p(g) = \mathrm{card}(\{g': f(g') \geq f(g)\})\ /m,$ where card(.) is the number of elements in the set (.) and m is the same number in the domain of the function f. As the case study for such analysis we selected 13 COGs of the 132 COGs studied in (V'yugin *et al.*, 2003), which possessed extreme values of $<R_g>$ (and for which $p(g) < 0.1$). A fragment of this set is givenin the upper part of Table 1. In an analogous manner we analyzed the rest of COGs.

Table 1. Fragments of the COG list sorted by the $<R_g>$ value, where $<<R_g>> = 1.4623$; $<<dF_g>> = 0.6985$

COG	$<R_g>$	p-value for $<R_g>$	$<dF_g>$ in %	p-value for $<dF_g>$
COG0351	2.57	0.0076	-1.6399	0.023
COG0171	2.26	0.015	-0.62327	0.16
COG0547	2.19	0.023	0.10913	0.39
COG0169	2.14	0.015	0.66859	0.58
COG0573	2.11	0.038	0.50343	0.52
COG0135	2.1	0.045	-1.2455	0.076
COG0581	2.03	0.053	-0.42877	0.23
COG0221	1.95	0.061	-0.52252	0.2
COG0159	1.93	0.068	-1.2074	0.083
COG0597	1.92	0.076	1.1511	0.69
COG0340	1.9	0.083	-0.47401	0.21
COG0105	1.89	0.091	0.20002	0.4
COG1488	1.89	0.098	0.92679	0.61
....	....	....	....	.....
COG0060	1.5	0.3	1.1761	0.7
COG0012	1.47	0.39	0.40489	0.48
COG0016	1.38	0.58	1.363	0.73
COG0049	1.29	0.77	0.099124	0.37
COG0048	1.25	0.83	-0.039805	0.34
COG0051	1.25	0.84	0.20717	0.41
COG0052	1.22	0.86	0.75545	0.58
COG0013	1.19	0.92	0.98059	0.64

A computer program selects 247 genes g from the remaining 109 COGs, for which $p(g) < 0.1$ for p-values defined for any of the two above defined statistics. We refer to these genes as *gained genes*. The gained genes are considered candidates for horizontal transfers and other gene gain events. We further consider the following two options. First, we assume that there are no gain events and calculate numbers of gene duplications, losses and gains (separately for each COG). In Table 2 we give the total (over 109 COGs) of these numbers (duplication, loss and gain separately). Secondly, all the genes identified as gain events by our approach were excluded from the domain of the corresponding tree mapping α and the same total estimates were calculated (the first case is called **non-GAIN scenario**, the second is **GAIN scenario**).

Table 2. Total number of duplications in groups of species

Group of species	non-GAIN scenario	GAIN scenario
Archaea	149	143
Gram-positive bacteria	54	55
Alpha-proteobacteria	7	7
Gamma&Beta-proteobacteria	207	202
Epsilon-proteobacteria	0	0
Clamydia&Spirochaetes	2	2
DMS	5	4
Thermotoga&Aquifex	0	0

Massive gene duplication attributed to the root of a phylogenetic group could be interpreted as a result of possible "*genome duplication*". Such is the set of 92 gene duplications assigned to the root of the subtree of Archaea. Another large group of 83 gene duplications was found in the gamma-proteobacteria group and assigned to the root of the species subtree (((*Pmu,Hin*),(*Eco,Buc*)),*Vch*). It also might result from ancient genome duplication (see Fig.). Massive gene duplication in the ancestor of *Vibrio cholerae* was independently postulated in (Heidelberg *et al.*, 2000).

Conclusion

A mathematical model of gene duplication and loss was applied to compute numerical characteristics measuring discrepancy between the gene trees and the species tree. Using integral characteristics of COG quality, we excluded a set of gene trees from the analysis. We also conclude that the total number of gene duplications assigned to internal nodes of phylogenetic groups are almost independent of the scenario chosen, GAIN r non-GAIN.

References

Mirkin B.G., Muchnik I., Smith T. A biologically consistent model for comparing molecular phylogenies // J. Comput. Biol. 1995. V. 2. P. 493–507.

V'yugin V.V., Gelfand M.S., Lyubetsky V.A. Identification of horizontal gene transfer from phylogenetic gene trees // Mol. Biol. 2003. V. 37(4). P. 571–584. (In Russian).

Heidelberg J.F. *et al.* DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae // Nature. 2000. V. 406. P. 477–483.

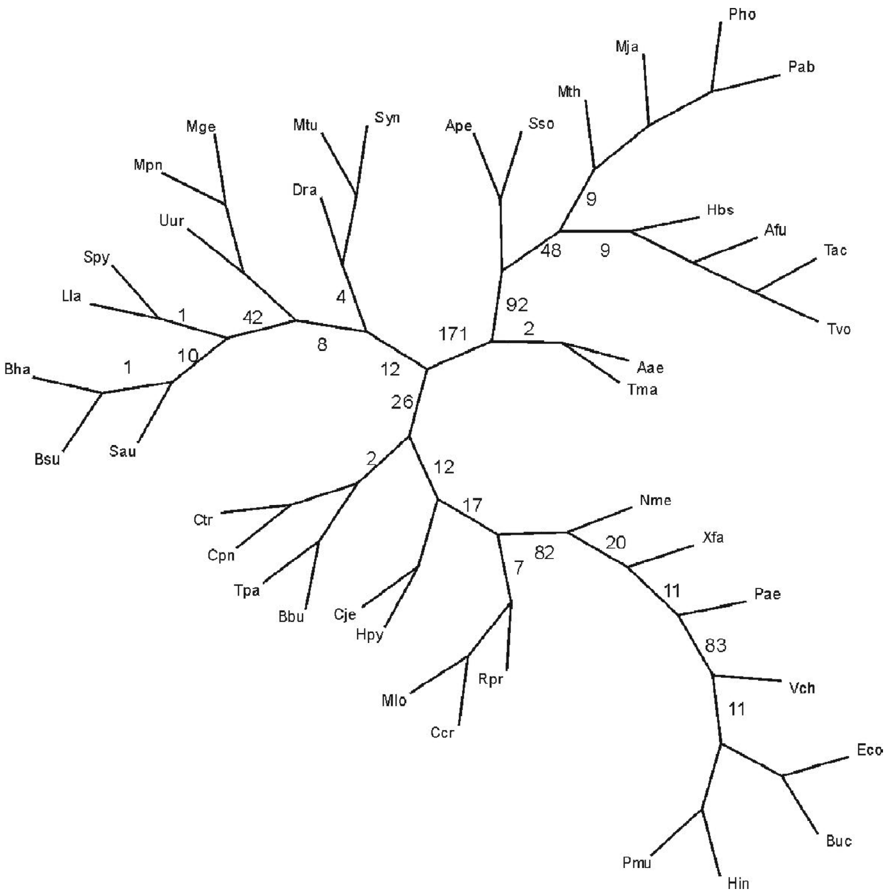


Fig. Total number of duplications assigned to groups of species (for the non-GAIN scenario).

DETERMINATION OF THE DEVELOPMENTAL AGE OF A *DROSOPHILA* EMBRYO FROM CONFOCAL IMAGES OF ITS SEGMENTATION GENE EXPRESSION PATTERNS

Myasnikova E.M.* [1], ***Reinitz J.*** [2]

[1] St.Petersburg State Polytechnical University, St.Petersburg, Russia; [2] University at Stony Brook, New York, USA
* Corresponding author: e-mail: myasnikova@spbcas.ru

Keywords: *gene expression, embryo staging, Drosophila, support vector regression*

Summary

Motivation: We address the problem of the temporal characterization of *Drosophila* embryos from confocal images of their segmentation gene expression patterns.

Results: We have developed a method for automated staging of an embryo on the basis of confocal images of a segmentation gene expression pattern. Phases of spectral Fourier coefficients were used as the features characterizing temporal changes in expression patterns. The age detection is implemented by applying support vector regression which is a statistical method for creating regression functions of arbitrary type from a set of training data. The training set is composed of embryos for which the precise developmental age was determined by measuring the degree of membrane invagination. Testing the quality of regression on the training set showed good prediction accuracy.

Availability: http://www.urchin.spbcas.ru/flyex/, http://flyex.ams.sunysb.edu/flyex

Introduction

Like all other insects, the body of the fruit fly *Drosophila* is made up of repeated units called segments. The genetic network which controls segmentation in *Drosophila* is one of the few genetic networks fully characterized genetically (Nusslein-Volhard *et al.*, 1984). The initial determination of segments is a consequence of the expression of 16 genes which are expressed in patterns that become more spatially refined over time. To provide full spatiotemporal information about expression of these genes the data must be obtained at cellular resolution in space, and at temporal resolution that is close to the characteristic time for changes in gene expression. In our experiments such data are obtained by means of immunofluorescence histochemistry and confocal scanning microscopy.

The problem of temporal characterization arises as gene expression data is acquired from fixed embryos, for which a precise developmental time is not known and thus the temporal dynamics must be reconstructed from many samples, each at a different stage of development. A fundamental step in such reconstruction is to determine the developmental age of each embryo. We consider embryos belonging to the cleavage cycle 14A, which lasts from 130 to 180 minutes after fertilization and egg deposition. At this time the segments are determined and the invagination of membranes and the cellularization of cells happens (Foe *et al.*, 1983). The method can be subdivided into two major stages, of which the first is the extraction of characteristic expression features of embryos of different age, and the second is standardization against morphological data, obtained from the independent source, for example, *in vivo* measurements of the degree of membrane invagination (Merrill, 1988).

Methods

Dataset. Our dataset contains confocal scans of 809 wild type embryos belonging to the cycle 14A. Of these, all are stained for the pair-rule segmentation gene *even-skipped* (*eve*) and two other genes that vary among the dataset. 120 of embryos from the dataset were rephotographed to visualize the morphology of the blastoderm, and the precise developmental age was determined for each of these embryos. The measured ages are distributed uniformly over the range from 20 to 60 minutes from the onset of cleavage cycle 14A. The 120 standardized embryos are used as a training set for temporal analysis.

Extraction of characteristic features. To detect the age of an embryo on the basis of knowledge about its *eve* expression pattern it is necessary to present the pattern in terms of a small number of parameters which well characterize temporal changes in *eve* expression domains. It has been shown in our previous study (Aizenberg *et al.*, 2002) that the frequency domain representation of images may be used to detect the characteristic features, which mark the development of expression patterns over time. The Fourier spectrum is extracted from two-dimensional images by means of the Fast Fourier Transform, and the phases of low frequency coefficients of Fourier spectra are considered as independent variables for the regression algorithm.

However, the spectral phases cannot be directly involved into the regression analysis for two reasons: first, phases are periodic values and, second, number of independent parameters is too big as compared with the size of the training set. To get rid of periodicity for each parameter the standard range of values is defined so that the maximal in absolute value pair-wise difference between values, which the parameter takes over the training set, is set to minimum. The problem of high dimensionality of the feature space often arises in the regression prediction. As spectral phases are strongly correlated and hence the feature set is redundant, it is possible to reduce the dimension of the feature vector by means of the principal component analysis (PCA).

Construction of the regression function. Each embryo of the training set is now characterized by a multidimensional vector containing as components the value of developmental age together with parameters of gene expression patterns. The regression function for the age prediction is created from the set of the training data applying the support vector regression (SVR) (Smola, 1998). SVR is a statistical method in pattern recognition theory, which is more flexible compared to classic regression because it allows for the use of loss functions of various types. The SVR algorithm as applied to embryo age detection from quantitative data on gene expression is described in more detail in (Myasnikova *et al.*, 2002), here we will give just the brief statement of the main ideas of the method.

Suppose we are given training data presented by L observations (embryos). Each observation consists of a pair: a vector of N characteristic features $x_i = (x_{1i},\mathrm{K}\ ,x_{iN})\quad i = 1,\mathrm{K}\ ,L$ and the associated 'truth' y_i (an embryo age), given us by a trusted source (membrane invagination). In linear e-SVR (Vapnik, 1995) the goal is to find a function $f(x) = (w, x) + b.$ that minimizes the regularized empirical risk functional

$$R_{reg}[f] = R_{emp}[f] + \frac{1}{2}\|w\|^2 = \frac{1}{L}\sum_{i=1}^{L} |\, y_i - f(x_i)\,|_\varepsilon + \frac{1}{2}\|w\|^2 , \tag{1}$$

where R_{emp} is the empirical risk with *e*-insensitive loss function given by

$|\xi|_\varepsilon = \begin{cases} 0 & if\ |\xi| \le \varepsilon \\ |\xi| - \varepsilon & otherwise \end{cases}$, $\frac{1}{2}\|w\|^2$ is a regularization term.

Age prediction. To determine the age of an embryo not belonging to the training set on the basis of a given confocal image of its *eve* expression pattern, the new image is subjected to the same preprocessing and feature extraction procedures as the members of the training set. Periodicity of the spectral phases is overcome by bringing their values to the standard range defined over the training set. Then applying the PCA the number of features is reduced to the required in SV regression number, and the embryo age is defined using the regression function constructed for the training set.

Results

The phases of low frequency Fourier coefficients up to the frequency 8 were extracted from the images which all were brought to the size 256 x 256 pixels. The total number of parameters extracted in such a way were 84, but due to the symmetry of the spectrum only 42 were pair-wise different. Applying the PCA we obtained 4 significant uncorrelated variables, which contained 96 % of the information originally contained in the whole set of 42 parameters. The regression function $f(x)$ was constructed from the training set of 120 embryos for which the precise developmental age was experimentally determined. The results of regression estimation are presented in the Fig. 1. The quality of regression is characterized by the minimal value of the cost function (1) with no regularization term, R_{emp}, i.e. the average *e*-insensitive deviation between observed and predicted ages. For our training set this value is achieved as 2.2 minutes.

To test the accuracy of prediction we consequently exclude, one by one, a single item from the training set, and use all the rest as a working set, thus predicting the age of the excluded embryo. As a criterion of the quality of prediction the risk function R_{emp} is used with the entries computed for the excluded items $\frac{1}{L}\sum_{i=1}^{L}|\, y_i - f_i(x_i)\,|_\varepsilon$, f_i here are the different functions each time newly estimated for the training set with the exclusion of *i*th embryo. The value of the criterion is equal to 2.4, and the results of testing are visualized in the Fig. 2.

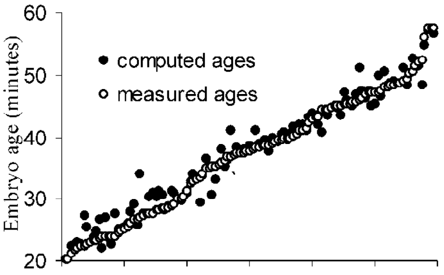


Fig. 1. Embryo ages (measured in minutes from the onset of cycle 14A) computed for the training set using the regression function (black); and measured in experiment (white).

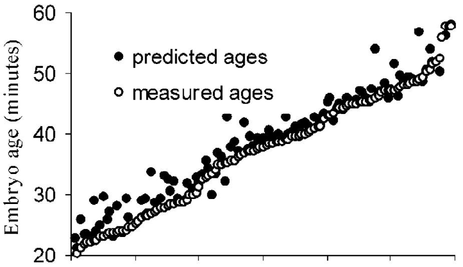


Fig. 2. Embryo ages predicted for the members of the training set excluded, one by one, from the regression analysis (black) and measured in experiment (white).

Discussion

In this paper we address the problem of temporal resolution of segmentation gene expression patterns by providing new methods for their temporal characterization. We have already reported the method for the detection of developmental age of embryos belonging to the late part of the cycle 14A (Myasnikova *et al.*, 2002). At that step of our study we restricted ourselves to the quantitative data extracted from the central 10 % horizontal strip running in an A-P direction on the midline of an embryo. The method allowed to predict the ages only to embryos in which the full set of seven *eve* stripes has been already formed, with the training set containing 103 embryos. Here we have extended our method to any confocal images of *eve* expression patterns presented in pixel format. The accuracy of prediction is almost of the same order as compared to the old method (2.4 vs 2.0 min.), while the method presented here, in contrast to the old one, doesn't require any preliminary manual work on extraction of characteristic features.

Acknowledgements

This work is supported by NIH grant RR07801 and GAP award RBO-1286.

References

Aizenberg I., Myasnikova E., Samsonova M., Reinitz J. Temporal classification of Drosophila segmentation gene expression patterns by the multi-valued neural recognition method // Mathematical Biosciences. 2002. V. 176. P. 145–159.

Foe V., Alberts B. Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in *Drosophila* embryogenesis // J. of Cell Science. 1983. V. 61. P. 31–70.

Merrill P., Sweeton D., Wieschaus E. Requirements for autosomal gene activity during precellular stages of *Drosophila melanogaster* // Development. 1988. V. 104. P. 495–509.

Myasnikova E., Samsonova A., Samsonova M., Reinitz J. Support vector regression applied to the determination of the developmental age of a *Drosophila* embryo from its segmentation gene expression patterns // Bioinformatics. 2002. V. 18. S87-S95.

Nusslein-Volhard C., Wieschaus E., Kluding H. Mutations affecting the pattern of the larval cuticle in *Drosophila melanogaster*. I. Zygotic loci on the second chromosome // Roux's Archives of Developmental Biology. 1984. V. 193. P. 267–282.

Smola A., Scholkopf B. A tutorial on support vector regression // NeuroCOLT2 Technical Report Series. NC2-TR-1998-030, http://www.neurocolt.com. 1998.

Vapnik V. The Nature of Statistical Learning Theory. Springer, N.Y. 1995.

YASS: ENHANCING THE SENSITIVITY OF DNA SIMILARITY SEARCH

*Noe L., Kucherov G.**

LORIA/INRIA-Lorraine, 615, rue du Jardin Botanique 54602 Villers-les-Nancy, France
* Corresponding author: e-mail: Gregory.Kucherov@loria.fr

Keywords: *local alignment, hit criterion, transition-constrained spaced seed*

Summary

We present YASS – a new tool for computing local similarities in DNA sequences. Similar to many existent algorithms (e.g. BLAST), YASS first searches for small patterns ("seeds") shared by input sequences and then extends some of those ("hits") into larger alignments. New features of YASS include, on the one hand, a new *transition-constrained seed model* and, on the other hand, a flexible *group criterion* of forming hits out of individual seeds. Computer experiments confirm that the program achieves a good sensitivity/selectivity trade-off compared to existing programs, which makes of YASS an efficient tool for comparing long DNA sequences, such as eukaryotic chromosomes.

Motivation

Sequence alignment is a Swiss-army knife for Bioinformatics, as most of comparative analysis relies on results provided by alignment programs. The well-known Smith-Waterman algorithm provides an exact algorithmic solution to the problem of computing optimal local alignments. However, its quadratic time complexity has been a motivation for the creation of rapid heuristic local alignment tools. FASTA (Lipman *et al.*, 1988), BLAST (Altshul *et al.*, 1990,1997), BLAT (Kent, 2002), PatternHunter (Ma *et al.*, 2002) BLASTZ (Schwartz *et al.*, 2003) are examples of such tools, to name a few. All those methods follow a seed-based approach for detecting alignments of interest. They first search for small patterns ("seeds") shared by input sequences and then extend some of those ("hits") into larger alignments of similarity regions. The central problem here is to increase the detection capacity (*sensitivity*) of the algorithm without sacrificing its *selectivity* (i.e. without increasing the number of spurious hits) that directly affects the time efficiency. This goal led to the development of new efficient *hit criteria*, i.e. criteria that define which patterns shared by two sequences are assumed to witness a potential alignment. Two types of improvements can be distinguished. On the one hand, using two or more closely located smaller seeds instead of one larger seed has been shown to improve the sensitivity/selectivity trade-off, especially for detecting long similarities. On the other hand, new seed models have been proposed, such as seeds with errors (Kent, 2002), spaced seeds (Ma *et al.*, 2002) or vector seeds (Brejova *et al.*, 2003). We propose a tool that brings further improvements in both directions above.

Method

We developed YASS (*Yet Another Similarity Searcher*) that implements an efficient and flexible local alignment strategy, achieving a good trade-off between sensitivity and selectivity.

Seed model

Seeds are specified by a seed pattern built over a three-letter alphabet {#,@,_}, where # stands for a nucleotide match, _ for a don't care symbol, and @ for a match or a transition (mutation A↔G or C↔T). The *weight* of a pattern is defined as the number of # plus half the number of @. The weight is a characteristic of seed selectivity.

The advantage of transition-constrained seeds over “regular” spaced seeds (seeds over the two-letter alphabet {#,_}) is based on the biological observation that transition mutations are relatively more frequent than transversions, both in coding and non-coding regions. Typically, biologically relevant alignments contain about the same number of transitions and transversions, while transitions are expected to be half less frequent for i.i.d. random sequences.

We have designed efficient transition-constrained seeds and measured their sensitivity using a Bernoulli model of DNA alignments (see also Kucherov *et al*., 2004). This analysis confirmed that transition-constrained seeds bring an improvement over spaced seeds provided the transition/transversion rate is over about 0.92, which is typically the case in real genomic sequences. Furthermore, we also studied the performance of transition-constrained seeds on a Markov model of order 5, that we constructed from a large mixed sample of about 100000 crossed alignments of genomic sequences of distantly related species.

By default, YASS uses the pattern #@#__##__#_##@# of weight 9 that provides a good compromise in detecting similarities of both coding and non-coding sequences. On the other hand, YASS allows the user to specify his own seed pattern. Several pre-selected patterns are provided by the YASS web interface.

Hit criterion

YASS is based on *a multi-seed* hit criterion which defines a hit as a group of closely located and possibly overlapping seeds. Two seeds belong to the same group if they occur within a bounded distance and, on the other hand, are located at close *dotplot* diagonals. Distance threshold parameters are computed according to probabilistic sequence models taking into account substitutions and *indels*, similarly to models used in (Benson, 1999). Note that seeds of a group are allowed to overlap. An additional *group size* parameter sets a lower bound on the total number of individual matches and transitions of the group. Using the group size results in a flexible criterion that combines a guaranteed selectivity with a good sensitivity both on short and long similarities. Further details can be found in (Noe, 2003).

Experimental Results

To illustrate the efficiency of YASS, we performed a series of comparative tests vs the bl2seq program from the NCBI BLAST package 2.2.6. Several complete bacterial genomes ranging from 3 to 5 Mb have been processed against each other using both programs. For each pair of sequences, we counted the number of alignments with E-value smaller than 10^{-6} found by each program. Furthermore, we counted, for each program, the number of those alignments detected exclusively by this program and not by the other. An alignment is considered to be detected by the other program if it contains, is contained in, or overlaps an alignment computed by the other program. To avoid a bias caused by splitting alignments into smaller ones we also computed the total length of exclusive alignments, found by each program.

A representative selection of results is given in Table. They imply that within a smaller execution time, YASS detects at least about 20 % more exclusive similarities that cover about twice the overall length of those found by bl2seq. Moreover, the gain in execution time increases considerably when the size of sequences gets larger. Other experiments have also been carried out to demonstrate the efficiency of transition constrained seeds compared with other models, on both coding and random sequences (http://www.loria.fr/projects/YASS/yass_experiments.html).

Table. Comparative tests of YASS vs bl2seq (NCBI BLAST 2.2.6). All tests use the scoring system +1/-1 for match/mismatch and -5/-1 for gap opening/extension. Only alignments with E-value < 10^{-6} are considered. Reported execution times have been obtained on a Pentium IV 2.4GHz computer

Sequence 1		Sequence 2		CPU time (sec)		# alignments		# exclusive align.		Excl. align. length	
				yass	bl2seq	yass	bl2seq	yass	bl2seq	yass	bl2seq
Synechocystis sp. PCC 6803	3.6 Mb	M. tuberculosis CDC1551	4.4 Mb	122	148	494	310	130	27	29145	7970
Synechocystis sp. PCC 6803	3.6 Mb	Yersinia pestis KIM	4.6 Mb	156	253	901	617	186	54	39354	19994
Synechocystis sp. PCC 6803	3.6 Mb	Vibrio p. RIMD 2210633 I	3.3 Mb	164	167	940	465	349	60	65788	28883
M. tuberculosis CDC1551	4.4 Mb	Yersinia pestis KIM	4.6 Mb	168	255	738	515	197	86	44348	23361
M. tuberculosis CDC1551	4.4 Mb	Vibrio p. RIMD 2210633 I	3.3 Mb	72	69	498	295	171	30	36474	12021
Yersinia pestis KIM	4.6 Mb	Vibrio p. RIMD 2210633 I	3.3 Mb	149	217	2900	1953	622	264	186585	110352

Software

YASS program is written in C and runs on Windows/Unix platforms (compiled using respectively MinGW and GCC. YASS is a free software and is distributed under the GPL Version 2.

A web server interface has been developed to query the program remotely, with the possibility to upload sequences (a small sequence database is available too), to set search parameters, to display a dot-plot or a list of output alignments, and to download result for post-processing. Server scripts have been developed in Perl-CGI and PHP and run on the Apache server.

Availability

A YASS stand-alone precompiled or source version can be freely downloaded from http://www.loria.fr/projects/YASS/ under the GNU General Public License. The YASS web server interface can be accessed from that page.

References

Altshul S., Gish W., Miller W., Myers E., Lipman D. Basic local alignment search tool // J. of Mol. Biol. 1990. V. 215. P. 403–410.

Altshul S., Madden T., Schaffer A., Zhang J., Zhang Z., Miller W., Lipman D. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs // Nucleic Acids Res. 1997. V. 25(17). P. 3389–3402.

Benson G. Tandem repeats finder: a program to analyse DNA sequences // Nucleic Acids Res. 1999. V. 27(2). P. 573–580.

Brejova B., Brown D., Vinar T. Vector seeds: an extension to spaced seeds allows substantial improvements in sensitivity and specificity // Algorithms in Bioinformatics. 2003. 2812 of LNBI, Springer. P. 39–54.

Kent W.J. BLAT-the BLAST-like alignment tool // Genome Res. 2002. V. 12(2002). P. 656–664.

Kucherov G., Noe L., Ponty Y. Estimating seed sensitivity on homogeneous alignments // Proc. of the IEEE 4th Symposium on Bioinformatics and Bioengineering (BIBE2004). 2004, Taichung,Taiwan, 2004. IEEE Computer Society Press.

Lipman D., Pearson W. Improved tools for biological sequence comparison // Proc. Natl Acad. Sci. 1988. V. 85. P. 2444–2448.

Ma B., Tromp J., Li M. PatternHunter: Faster and more sensitive homology search // Bioinformatics. 2002. V. 18(3). P. 440–445.

Noe L., Kucherov G. YASS: Similarity search in DNA sequences // Res. Report RR-4852, 2003. INRIA. http://www.inria.fr/rrrt/rr-4852.html

Schwartz S., Kent J., Smit A., Zhang Z., Baertsch R., Hardison R., Haussler D., Miller W. Human-mouse alignments with BLASTZ // Genome Res. 2003. V. 13. P. 103–107.

Smith T., Waterman M. Identification of common molecular subsequences // J. of Mol. Biol. 1981. V. 147. P. 195–197.

REVELATION AND CLASSIFICATION OF DINUCLEOTIDE PERIODICITY OF BACTERIAL GENOMES USING THE METHODS OF INFORMATION DECOMPOSITION AND MODIFIED PROFILE ANALYSIS

Shelenkov A.A.*, Chaley M.B., Korotkov E.V.

Center of Bioengineering, Moscow, Russian Academy of Sciences
* Corresponding author: e-mail: fallandar@mail333.com

Keywords: *latent dinucleotide periodicity, prokaryotic genomes, information decomposition, modified profile analysis*

Summary

Motivation: Information decomposition (ID) of symbolical sequences is a powerful tool for studying the periodicity of symbolical texts. We applied this approach to show the presence of many sequences in prokaryotic genomes that have the latent dinucleotide periodicity, but have not been found yet.

Results: We found more than 2500 DNA sequences having latent periodicity with period length equal to 2 bases. The classification made has shown that all of the periods found could be merged into 116 classes. The method of modified profile analysis (MPA) has revealed more than 2000 DNA sequences from bacterial genomes belonging to the classes found having the latent periodicity with insertions and deletions (IaD) of symbols. The functional and evolutional meaning of latent dinucleotide periodicity in different prokaryotic genomes is discussed.

Methods and Algorithms

We have shown earlier (Korotkov *et al.*, 2003) that the methods of finding periodicity in symbolical sequences based on the Fourier transformation and dynamic programming have a number of essential constraints that do not allow to reveal a feebly marked periodicity in symbolical sequences. We developed the method of ID for revealing a feebly marked or latent periodicity in symbolical sequences to avoid many disadvantages intrinsic to the methods based on the Fourier transformation or dynamic programming (Korotkov *et al.*, 2003). However, ID method in its present form does not allow to reveal a feebly marked periodicity with a presence of symbol IaD. That is why the aim of this work was to reveal all dinucleotide periods from the known genomic sequences of prokaryotes, to classify them and then to apply the MPA for the search of latent periods with IaD of symbols.

Implementation and Results

We have applied the ID method to search the latent periods with length equal to 2 bases in different DNA sequences from the bacterial genomes. Using this method we have revealed more than 2500 DNA sequences from the latest version of Genbank having different types of dinucleotide latent periodicity. The computational complexity of this problem is comparatively high, so we used parallel cluster calculations. For being able to perform such calculations, we have transformed our software so that it could use the MPI subroutines for the data transmission between individual processors. Then we have classified the periods found by using the information metric to determine the homology of periodicity matrices. For doing this, all matrices were normalized to the same sum of their elements. We then performed their pairwise comparison taking into account all their cyclic shifts and complementary transformations. The two most homologous matrices were merged and then the pairwise comparison was performed again. This classification let us found that there are

116 types of dinucleotide periodicity in prokaryotic genomes. The example of the periodicity class (or type) revealed is shown in Table 1.

We used the class matrices obtained in the method of MPA to search for the latent periodicity with IaD (Chaley *et al.*, 2003; Korotkov *et al.*, 2000). We took a 200–300 symbol window and scanned all known DNA sequences from prokaryotic genomes with it. Random texts having tenfold length were analyzed to control the chosen level of statistical significance. We have not found the latent periodicity specified by the class matrices in these texts. We have found totally more than 2000 new sequence in which the latent periodicity has been presented with IaD.

Table 1. The 41st class of periodicity with the length of 2 DNA bases. The mean popularity of each base type for 2 period positions is shown

	1	2
A	9.9	25.5
T	27.2	11.0
C	2.0	0.0
G	1.0	4.0

It is interesting that the Fourier transformation method does not reveal the latent periodicity found by the ID method. The reason is the rather large lengths of found periods, compared with the size of the alphabet used and the low level of similarity between periods. The power of a long period, as noted above, is distributed on powers of a set of short periods, leading to difficulties in the detection of the latent periodicity. We have also shown (Korotkov *et al.*, 2003) that the method of tandem repeat finding based on dynamic programming did not reveal the periods found by our method.

Let us consider some examples of latent periodicity with IaD for some class of matrices (period length is equal to 2 symbols). In Table 1 the matrix of base popularity for the 41st of 116 total classes is shown. First row shows the period positions and first column shows the DNA bases. The 41st class is remarkable in a sense that its appearance with insertions and deletions in prokaryotic genomes is the most wide-spread.

In Table 2 the sequences themselves are shown. They are aligned using the matrix from Table 1.

K1 – the beginning of the latent periodicity sequence in 200-symbols window; K2 – the end of the latent periodicity sequence in 200-symbols window; N1 – the beginning of the latent periodicity sequence in the sequence analyzed; N2 – the end of the latent periodicity sequence in the sequence analyzed; Z – statistical significance of the found periodicity, used as score. Two sequences one under another are shown in Table 2. The upper sequence is the sequence from the Genbank locus that has been analyzed. Lower sequence is the consensus sequence of the 41st latent periodicity class.

We can see from Table 2 that the sequences found are aligned to the profile used with a small number of IaD.

In Fig. 1 one can see that all of the shown examples of latent dinucleotide periodicity regions belong to the non-coding regions of genomes. Most likely that these sequences are the ancient microsatellite sequences that have diverged greatly and have accumulated a significant number of base IaD. This fact did not allow identify these sequences as microsatellite ones by means of standard approaches and only the ID method together with dynamic programming is able to do this.

From the Fig. 2 it can be seen that the ID in its present form can not be used to reveal the latent periodicity with IaD. It becomes able to reveal the periodicity on the statistically significant level only after performing the alignment to the 41st class of DNA sequence dinucleotide periodicity. This comparison shows that the classification of the dinucleotide latent periodicity of symbolical sequences types performed by us allows to reveal yet unfound sequences having latent periodicity of microsatellite type. Revealing of such sequences is very important for performing different types of PCR analyses and the development of polymorphous markers for the quick diagnostics of different genetic factors. In particular, using this approach has allowed developing the polymorphous markers for distinguishing bacterial strains of the same type in Center of Bioengineering of RAS.

Table 2. Four examples of DNA base sequences from the bacterial genomes having the latent periodicity with the length of 2 bases belonging to the 41st class with IaD

Locus AB005787 6801 bp DNA BCT 27-JUL-1997 ACCESSION AB005787 K1= 88 K2= 199 N1= 5633 N2= 5742 Z= 7.6 atatattt-taaatat-tagctactccat-taccgatatttacctaaaaataagttttttacat-taatatatatatatatatatata-tatatatatatatatatata tatatatatatatatatatatatatagatatttccttctgactatatatagctactatacatacattatatatatatatatatatatatatatata-tatatatat
Locus AB012620 1005 bp DNA BCT 26-MAR-1999 ACCESSION AB012620 K2= 200 K1= 52 N1= 122 N2= 267 Z= 7.8 tctatttatatattcatataca-atatcactatttgcctccatgt-tctttataaacatacttaga-aatatatatatatatatatatatatatatatatatatatat atatatatatatatatatat ta-atat-tatctctcaaagatatatatggtttactagctttagtaagttaaagatataaacggttctatatatatatata-tatatatata-tatat-atat atatatatatatatatatatatatatatcatatctatacctattatatatatatatata
Locus AB013913 247 bp DNA BCT 01-DEC-1999 ACCESSION AB013913 K2= 200 K1= 52 N1= 73 N2= 209 Z= 7.3 taaatgtttatat-tatttgaata-aaacattcaaata-atata-aaaaata-atatatatattgacaatatatatatatatatatatatatatatatatat atatatatatatatatatatatatatattggat-taaacaaagatatatat-tat-tctatgt-tgtatgaacaaat-tggcaaaatagagatggaatata tatatatatatatatatatatatatatatatatatatatatatatatatatatatatat-agataaaaatatatatatatata
Locus AB014075 14043 bp DNA BCT 08-MAY-1999 ACCESSION AB014075 K2= 199 K1= 3 N1= 13151 N2= 13346 Z= 7.9 Atatagggtatctttttctttata-ataggtatagcta-atcaataacaata-atacatattgaattatatatat-atatatatatatatatatatatatat atatatatatatatatatatatatatatatcttagccataaatgataaagtatagtatagtcctttatttttgaggtgatata-atgtatgaatatatatat atatatata-tatat-atat-atatatatatatatata-tata-tatatatatatatatatataaaa-aggtacatatatgggcataaacaaggagtatata gttatcgaaa-atggagatat-tggatatatatatatatatatatatatatatatatat-atatatatatatatatatatatatatatatat

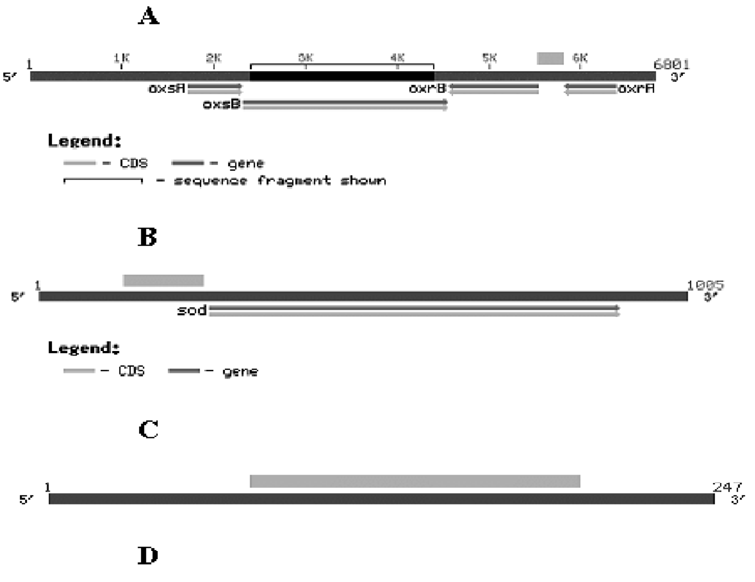


Fig. 1. The schemes of sequences (loci from Genbank) in which the latent periodicity of 41st class with IaD have been revealed. The corresponding DNA sequences and the obtained alignment to the weight matrix of 41st class are shown in Table 2. The graphs of ID are shown in Fig. 2. A – locus AB005787; B – locus AB012620; C – locus AB013913; D – locus AB014075. The latent periodicity regions of 41st class with IaD location are shown above the sequences.

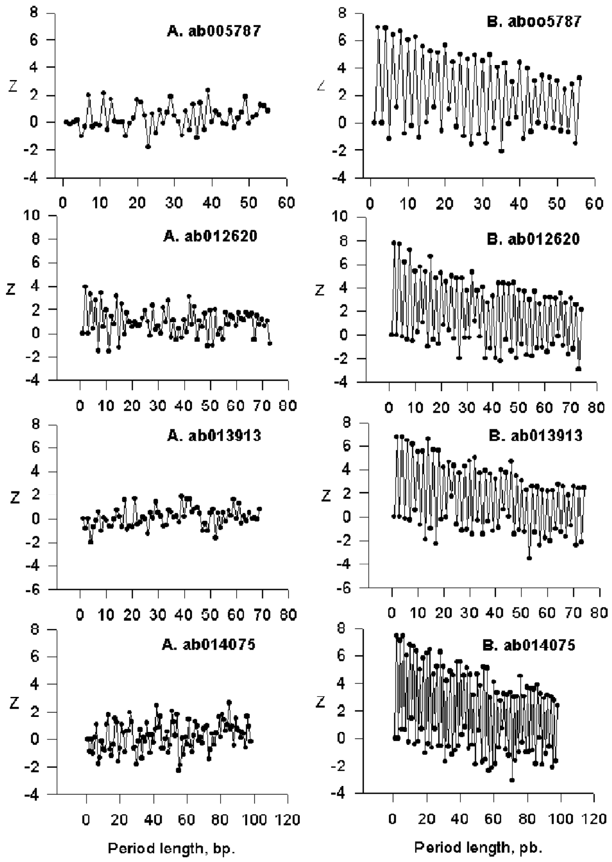


Fig. 2. The ID of sequences shown in Table 2. Graphs on the left (A) show the ID of sequences without IaD. Graphs on the right (B) show the sequences ID after performing the alignment to the matrix of 41st class of periodicity.

References

Korotkov E.V., Korotkova M.A., Kudryashov N.A. Information decomposition method to analyze symbolical sequences // Phys. Let. A. 2003. V. 312. P. 198–210.

Chaley M.B., Korotkov E.V., Kudryashov N.A. Latent periodicity of 21 bases typical for MCP II gene is widely present in various bacterial genes // DNA Sequence. 2003. V. 14. P. 37–52.

Korotkov E.V., Korotkova M.A., Rudenko V.M. MIR: The family of repeats common for the genomes of many vertebrates // Mol. Biol. 2000. V. 34. P. 553–559.

COMPUTER ANALYSIS OF MULTIPLE REPEATS IN BACTERIA

Vitreschak A., Noe L., Kucherov G.*

INRIA-Lorraine/LORIA, 615, rue du Jardin Botanique, BP 101, 54602 Villers-lès-Nancy, France
* Corresponding author: e-mail: vitresch@loria.fr

Keywords: *repeats, clusterization, bacteria*

Summary

Motivation: The presence of repeated sequences is a well-known feature of bacterial genomes and interpretation and classification of those repeats is an actual problem.

Results: We described a method for computing *multiple repeats*, that is sequences that have multiple (two or more) occurrences in a genome. In order to identify multiple repeats in bacteria genomes, we apply the YASS software (Noe, Kucherov, 2004) and developed a novel algorithm for multiple repeat clusterization. Exhaustive computation and analysis of those "clusters of repeated sequences" in bacteria is the subject of the present work.

Availability: Program is available by e-mail: vitresch@loria.fr

Introduction

The presence of repeated sequences is a well-known feature of bacterial genomes. In general, a DNA repeat is a sequence, which appears at least in two copies in the genome. The size of repeated sequences and their biological function differ greatly: in one case, a repeat can be about a thousand nucleotides long and contain coding open reading frames (for example, a mobile element); in other cases, a repeat can correspond to a regulatory element located in intergenic regions. Moreover, repeated sequences can be strongly conserved not only within one genome, but also across different (in some cases remotely related) genomes.

There are several programs specially devoted to the computation of repeats within a given genomic sequence (Kurtz *et al.*, 2001; Vincens *et al.*, 1998; Lefebvre *et al.*, 2003). Alternatively, such repeats can be obtained by computing, using any local alignment method, local similarities between the input sequence and itself. On the other hand, there is no method to systematically compute *multiple repeats*, that is sequences that have multiple (two or more) occurrences in a genome. Exhaustive computation and analysis of those "clusters of repeated sequences" in bacteria is the subject of the present work.

Data and Methods

In order to identify multiple repeats in a bacterial genome, we first apply the YASS software (Noe, Kucherov, 2004) and find all strong local similarities, viewed as two-copy repeats, within the genome. YASS parameters have are set to detect 70 % similarity alignments with a very low false positive rate (using the seed ##@_#@_#_#__## of weight 8 and group size 11, for details see Noe, Kucherov, 2004). All possible repeated sequences found by YASS are then grouped into clusters, with the goal that each cluster contains all copies of the same repeated biological element.

The clusterization of possible repeats is made in two steps. The first step (pre-clustering) consists in processing all local alignments found by YASS. This pre-clustering step groups together sequences that are strongly related: this is achieved by a heuristical search for quasi-cliques (almost perfect cliques) in the graph in which nodes are sequences and edges are similarities. The data structure used at this step is an interval tree that stores the coordinates of each sequence occurring in each YASS alignment. These initial clusters are "starting points" for further clusterization and are essential for the stability of "cores" of clusters (see below).

A method of "cores" is used at the second step of clusterization. Its main idea consists in using most conserved parts of repeats, called "cores", for controlling the clusterization process. First, a graph is constructed with nodes corresponding to the initial clusters. An edge connects two nodes when at least one sequence from one initial cluster "overlaps" at least one sequence from another initial cluster. Additional conditions for connecting two nodes (initial clusters) are the following:

$$\min\,(L_1/L_{overlap},\, L_2/L_{overlap}) < 2\,, \qquad (1)$$

$$\max\,(L_1/L_2,\, L_2/L_1) < 2, \qquad (2)$$

where $L_{1,}$ L_1 and $L_{overlap}$ are the lengths of first repeat, second repeat and the length of common part (overlap), respectively.

The first rule means that the length of the common part is at least a half of the minimal length of the two repeats. The second condition insures that the two involved sequences have comparable lengths.

Detected repeats from initial clusters correspond either to an entire repeated element or only to its part. In some cases, repeated elements correspond to a superposition of two or more different adjacent repeats (sometimes partially overlapped). This is an additional difficulty for the appropriate detection of repeated units. For example, if only rules (1), (2) are used for clusterization, then the process can result in a huge cluster containing more than 95 % of initial clusters (as applied to the *Neisseria meningitidis* genome). This fact is due to adjacent locations of distinct repeated elements on the DNA sequence, that can erroneously fall into one cluster. A simple illustration is given in Figure 1. The initial cluster1 is joined with the initial cluster2 and the latter is joined with the initial cluster3. cluster2 contains parts of both repeat1 and repeat2 and because of this "bridge", initial clusters 1 and 2 are also joined. In this way, different non-related repeats can be joined together, and the whole process results in one huge "supercluster". To cope with this problem, a method of "cores" has been developed.

At the second step, a "core interval" (core) is computed for each cluster. The core corresponds to the most conserved part of the repeat and core coordinates are computed as the average of corresponding sequence coordinates of the cluster.

Using the cores, the clusterization step is defined as the following traversal of the set of clusters (Fig. 1A): (a) after constructing the set of initial clusters, choose a start initial cluster (the largest one) (b) iteratively join the current cluster with other clusters which verify rules (1), (2) *applied to cores*. Manipulating cores allows us to avoid joining unrelated clusters, as shown in Figure 1. Figure 2B illustrates that those clusters are not joined anymore since rules (1), (2) are not verified for cores.

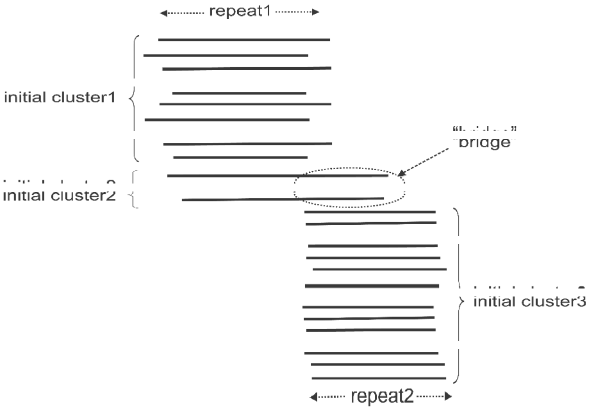


Fig. 1. Using the cores, the clusterization step is defined as the following traversal of the set of clusters (Fig. 1A): (a) after constructing the set of initial clusters, choose a start initial cluster (the largest one) (b) iteratively join the current cluster with other clusters which verify rules (1), (2) *applied to cores*. Manipulating cores allows us to avoid joining unrelated clusters, as shown in Figure 1. Figure 2B illustrates that those clusters are not joined anymore since rules (1), (2) are not verified for cores.

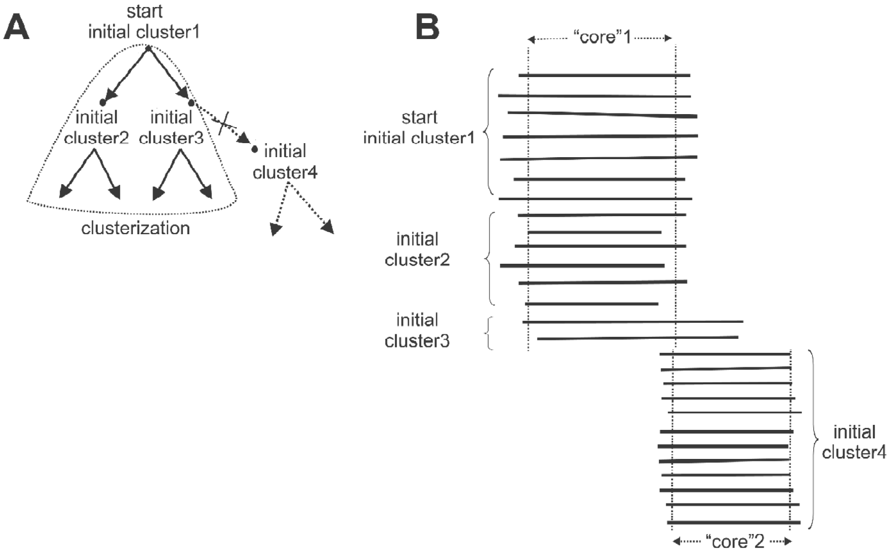


Fig. 2.

Results

We run our method on the *Neisseria meningitidis* genome and obtained a number of interesting clustered repeats, some of them with a known well-identified biological function. Interestingly, one resulting cluster embraced several hundreds of ρ-independent terminators. Several other clusters corresponded to mobile IS-elements (IS30, IS1016C2, IS1106).

Besides of those known elements, some interesting unknown repeats have been detected. For example, we found a cluster of sequences of about 120 bp long, which are highly distributed in the genome (more than 100 copies). These repeated sequence has a complex palindromic structure and is located in intergenic regions only, which suggests its possible regulatory role. Alternatively, this repeated element might be an RNA with a strong secondary structure, or a short mobile element of a new kind (suggested by its high degree of distribution). Another complex repeated element, revealed by our procedure, are also located in non-coding regulatory regions often adjacent to genes involved in bacterial pathogenesis. This demonstrates that the proposed clusterization method allows us to detect new repeats with unknown biological function. Interpretation and classification of those repeats is the subject of our current work. Program is available by e-mail:vitresch@loria.fr.

References

Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale // Nucleic Acids Res. 2001. V. 29. P. 4633–4642.

Lefebvre A., Lecroq T., Dauchel H., Alexandre J. FORRepeats: detects repeats on entire chromosomes and between genomes // J. Bioinformatics. 2003. V. 19. P. 319–326.

Noe L., Kucherov G. YASS: enhancing of sensitivity of DNA similarity search. submitted to BGRS-2004.

Noe L., Kucherov G. YASS: enhancing of sensitivity of DNA similarity search // Research report RR-4852, INRIA. http://www.inria.fr/rrrt/rr-4852.html [In French]. 2004.

Vincens P., Buffat L., Andre C., Chevrolat J.P., Boisvieux J.F., Hazout S. A strategy for finding regions of similarity in complete genome sequences // Bioinformatics. 1998. V. 14. P. 715–725.

BGRS 2004

BIOINFORMATICS AND EDUCATION

BIOINFORMATICS TEACHING IN MOSCOW ENGENEERING PHYSICS INSTITUTE (STATE UNIVERSITY)

Korotkova M.A.

Moscow Engineering Physics Institute, Moscow, Russia, e-mail: bioinf@rumbler.ru

Keywords: *education, bioinformatics, Data bank, mathematical model, computer analysis*

Summary

Motivation: Now there are many Data Banks with biological and genetics information and the exponential growth of the quantity of this information. This growth leads to the necessity of personnel training for analysis of this information. Those specialists must be able work with Data banks as advanced user and develop own algorithms and programs for genome analysis.

Results: We propose the program for bioinformatics teaching and had graduate several groups of students. Some of this specialists work in institutes of Russian Academy of Sciences, publish articles in leading science journals.

Availability: http://cyber.mephi.ru/

Introduction

Realization of Genome Program, obtaining of Human genome and developing of International Data Banks containing genomes of different genus lead to necessity of training for work with this information. Specialists in bioinformatics must be able work with Data banks as advanced user and develop own algorithms and programs for genome analysis, for the constructing of models of biological objects. The goal of bioinformatics teaching in MEPhI is to train students for such investigations.

Some years ago bioinformatics teaching started in MEPhI. This specialization give applied mathematician qualification to the students and unites computer and biological curses.

Methods and Algorithms

The specialization require training for discrete math and continuous math, as knowledge in mathematical statistics also.

Cybernetics department carry out bioinformatics teaching in MEPhI as having the math and computer training most closed to that is needed for specialization.

Bioinformatics training is realized in cooperation with the Bioingeneering center of Russian Academy of Science, one of leading organization of Russia in this area. This cooperation allows to get leading scientists to take part in education. Those scientists give lectures in biological courses and in modern methods of genome analysis.

Great number of genome information and growing number of international genome Data banks leads to necessity to attend working and creation such data banks in students teaching. Corresponding courses include theoretical aspects about information containing in such data banks and practical work using Internet.

Bioinformation teaching in MEPhI consists of three main parts: basic university education, specialty subjects of cybernetics department, specialization subjects of bioinformatics.

Basic university education include general chemistry, mathematical analysis, analytical geometry, physics, linear algebra, differential equations, complex variable functions theory and several subjects of the humanities. This training gives to students the main knowledge necessary for study special subjects.

Specialty subjects of cybernetics department such as informatics basics, some mathematical subjects as discrete mathematics, probability theory, methods of optimization, information security, neural networks theory basics and technical subjects such as open systems correlations, modern computer architectures, network operating systems, basics of automated informational technology, database design, computer graphics and some others. Knowledge obtaining in this courses enable students work with modern software and build up data bases in different data domains.

Specialization subjects such as biological objects organization and heredity principles, biological systems organization, proteins structure and functions, protein sequence analysis, mathematical methods of symbolic sequences analysis, theory of molecular evolution, mathematical methods of populational genetics, biological data bases. Those subjects enable students work in bioinformatics and as use web possibilities of Internet programs of international data banks as create own algorithms and programs for genome analysis.

Implementation and Results

There have graduated several groups of students trained in bioinformatics. Those students successfully work in different areas of bioinformatics. Some of these specialists work in institutes of Russian Academy of Sciences, publish articles in leading science journals. There was prepared dissertation in bioinformatics by our post graduate student. Some students work in other areas of computer science as nobody can predict research and development perspectives of students.

So we can see that our graduate students may work in bioinformatics and other areas of computer science. This shows the success of our programs and teaching.

RESEARCH AND TEACHING AT THE COLOGNE UNIVERSITY BIOINFORMATICS CENTER (CUBIC) – MOLECULAR NETWORKS IN ORGANISMS

*Lohmann M.**, *Schomburg D.*

Cologne University Bioinformatics Center, Cologne, Germany
* Corresponding author: e-mail: mark.lohmann@uni-koeln.de

Keywords: *metabolome, simulation, database, flux balance, petri net, protein design, education*

Introduction

The Federal Ministry of Education and Research (BMBF) supports bioinformatics in Germany with its “Training and Technology Initiative” since September 2000. It is intended to combine the bioinformatics activities of the respective special disciplines in Germany under way at the most different levels, network them with other fields of knowledge and thus create the prerequisites for qualitatively new findings. In order to ensure efficient technology transfer, the incorporation of industry is essential. Furthermore, the collaborative projects should be established in close coordination with the respective federal state governments so as to incorporate the training of young scientists. This implies the short-term establishment of advanced interdisciplinary courses of study.

In the summer of 2001 an international advisory board has selected six BMBF-sponsored bioinformatics centers which merged to form the Network of Bioinformatics Competence Centers (NBCC; http://www.nbcc-online.de). The development of effective tools for the utilisation of the results of the genome projects in interdisciplinary working groups as well as the establishment of bioinformatical standards form the principal tasks of the network. Therefore, all data necessary for the understanding of gene functions are generated, collected and analysed. Another main focus lies in the education of highly qualified bioinformaticians. In cooperation with universities, industry and state governments the centers offer degrees or postgraduate courses in bioinformatics.

Overview about the research activities at CUBIC

Metabolome Research. Based on the superior topic “Molecular Networks in Organisms” the Cologne University Bioinformatics Center (CUBIC; http://www.cubic.uni-koeln.de) combines the long-term mission: “in silico simulation of a whole organism” with short and medium-term goals. The research projects at CUBIC are primarily focused on the parallel analysis of experimental genome, transcriptome, proteome, structure, function and metabolome data for the in silico simulation of biological processes. The development of experimental methods for the analysis of the metabolic profile of a procaryotic model organism by using GC-MS, knockout experiments and 2D- gelelectrophoresis is an essential part on the way to clarify the complete network with all its regulative interfaces. Up to now we are able to detect 320 metabolites simultaneously whereof 120 are chemical identified so far. The developed methods form the foundation for genotype-phenotype correlation in order to identify unknown gene functions.

The metabolome analysis leads to a large quantity of data which have to be sorted analysed and interpreted. Based on the BRENDA (http://www.brenda.uni-koeln.de) model we develop an integrated database containing function and property data of biomolecules and organisms.

Whereas sequence or structural data for proteins are either right or wrong in the experimental context, other molecular data depend much more strongly on details of the experiments, or, even worse, are described in way which does not allow a comparison between different laboratories, molecules, or organisms. Since the automatisation of the methods requires definitions for standardization and quality we also develop experimental standards and the application of a standardized ontology.

A more comprehensive picture of biochemical processes in a model organism can be obtained from the analysis of substrate-product chains that can be extracted e.g. from a database, preferably BRENDA. A reliable prediction of cellular processes requires the integration of simulation techniques on various levels, from quantum-chemical calculations for the calculation of charges, electrostatic potentials and chemical reactions via force-field calculations for structure optimization, heuristic calculations for protein structure prediction to docking calculations. Starting with flux balance analysis (FBA) we have a tool to get insight into a metabolic system assumed to consist under steady state conditions. This procedure shall be improved by implementing experimental data. Concerning metabolic concentrations and the laws of thermodynamics our approach should yield in results better approximating real-life systems than FBA alone.

In this context we also build up a high-level Petri Net and include as much information as is available for a selected metabolic network. This will include Gibbs energies and, on a next level, regulative elements.

Protein Design. Another main research project is concerned with the development and optimization of computer tools for potein drug screening and design. It contains the implementation of fast and precised docking algorithms as well as an energy function which parameterizes and verifies possible confirmations of protein-protein docking simulations.

The protein engineering process for designing efficient enzymes is still solely based on experience. There is no common automatic method to predict the effect of a mutation on the stability of a protein. The protein-design group at CUBIC developed a distance- and direction-dependent knowledge-based potential and evaluated it for the prediction of protein-thermostabilities.

A further research group works on loop -prediction which is an important sub-task of homology modelling. To increase the quality of knowledge based loop-prediction an improved anchor group positioning and more complete databases are essential. The group developed an approach for supplying a more complete database and implementation of an improved loop-prediction algorithm.

Tools for Automated Structure Elucidation of Biological Metabolites. The Independent Junior Research Group develops the SENECA system for automated structure elucidation and identification of metabolites based on spectroscopic data. It further maintains the NMRShiftDB database, a web-based open-access information system for organic chemical structures and their Nuclear Magnetic Resonance (NMR) spectra (http://www.nmrshiftdb.org).

Know-how Transfer

Since April 2002 CUBIC offers a one year postgraduate course in bioinformatics at the University of Cologne. In this course scientists, with educational background in life sciences, mathematics, physics or computer science, will learn skills and techniques in bioinformatics.

The content of the programme is given in two modules and a thesis. Depending on educational background students will be first taught either in basics of life sciences or in computer science and mathematics. After this introductory course which should facilitate the gateway into bioinformatics, the course continues with the education in applied bioinformatics.

In order to gain insight into current computational life science research, the course participants will submit a twenty-week thesis related to bioinformatical problems or techniques. This thesis can also be prepared in external research institutions or in industry.

Each student has an individual computer for full-time access for practice. To guarantee a first-rate education CUBIC limits access to 30 students per year. The center offers scholarships for highly qualified students. The scripts of lectures, exercises and additional information material are available at the CUBIC-Intranet. The course is designed for international participants and completely taught in English.

Author index

Keywords

Подготовлено к печати
в редакционно-издательском отделе ИЦиГ СО РАН

Подписано к печати 23.06.2004
Формат бумаги 70x108/16. Усл.-печ.л. 38,22. Уч.-изд.л. 32,2
Тираж 250. Заказ 335
Отпечатано в типографии СО РАН
630090, Новосибирск, Морской просп., 2